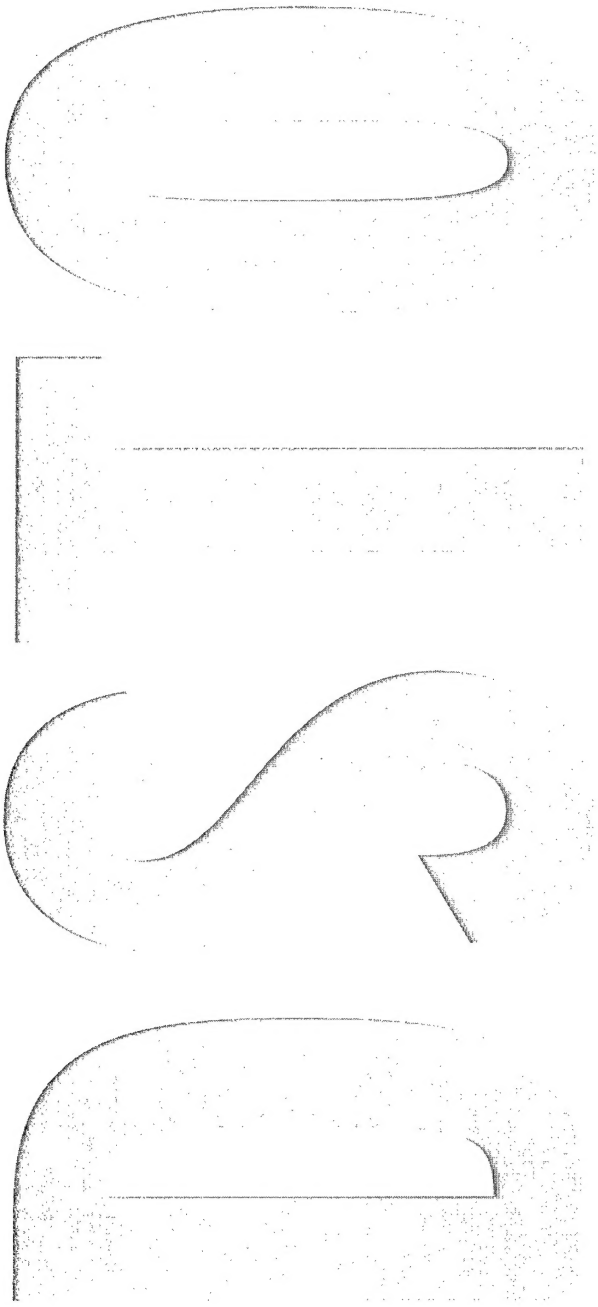
DEPARTMENT OF DEFENCE
DEFENCE SCIENCE & TECHNOLOGY ORGANISATION
DSTO



Effects of Noise on Almost Collinear Systems

Frank G. Polanco

DSTO-RR-0204

Effects of Noise on Almost Collinear Systems

Frank G. Polanco

Airframes and Engines Division
Aeronautical and Maritime Research Laboratory

DSTO–RR–0204

ABSTRACT

We investigate the effects of noise on developing predictions of ill-conditioned systems from measurements. In particular we investigate collinearity between measurement devices. We assume the system is linear in the measurements taken, and that the measurement noise is uncorrelated both to the true measurements and to other measurement noises. The "matrix" and "vector" techniques (two stress transfer function techniques developed earlier) are analysed. The matrix technique produces better results, but requires external system information during calibration. On the other hand, the vector technique (based on least squares) is easily implemented (no knowledge of external information is required), but is sensitive to ill-conditioned configurations of measuring devices. The vector technique is shown to be the well-known errors-in-variable model, and hence unbiased but inconsistent, which explains the large errors it produces. Although a correction to the vector technique improves results, it is still not as accurate as the matrix technique. This vector correction additionally requires estimates of noise in the measuring devices, and suffers from sensitivity to noise estimation errors. The surrogate-matrix technique substitutes internal for external system information, circumventing the need for external system measurements. Simulation results involving a simple truss support all theoretical findings. The surrogate-matrix and vector techniques are recommended for ill- and well-conditioned systems respectively.

Published by

DSTO Aeronautical and Maritime Research Laboratory
506 Lorimer St,
Fishermans Bend, Victoria, Australia 3207

Telephone: *(03) 9626 7000*
Facsimile: *(03) 9626 7999*

AR No. AR–011–785
March, 2001

APPROVED FOR PUBLIC RELEASE

Effects of Noise on Almost Collinear Systems

EXECUTIVE SUMMARY

Many aircraft components have a life span that is smaller than the design life of the aircraft. As such these components have to be replaced when their respective component retirement times (CRTs) are reached. The CRTs are calculated using the assumed loading spectrum together with experimentally derived component fatigue strength data. However, as the loading spectrum is assumed rather than being measured, the CRTs must be based on a worst-case spectrum and tend to be overly conservative, resulting in the retirement of most components well before their true "use-by-date". On the other hand, some components may be particularly sensitive to changes in the design loading spectrum. In such cases, any deviation of the actual load spectrum from the design spectrum may lead to the safe fatigue life being reached before the CRT. An accurate prediction of the loading spectrum, through actual measurements, has the potential not only to reduce operating costs, but also to increase safety.

Most fatigue life-limited components on helicopters are located on the rotor system and its drive train, and hence are dynamic in nature. However, due to the harsh working conditions on these rotor components, measurement instrumentation tends to be short lived. In addition, the transmission of measurements on rotor components to a storage device is itself non-trivial. One way to circumvent these direct measurement problems is to develop transfer functions between measurements on fixed helicopter components and measurements on dynamic (rotor) components. Knowing some set of parameter values measured on the fixed airframe then allows us to determine the loading on rotor components using these transfer functions.

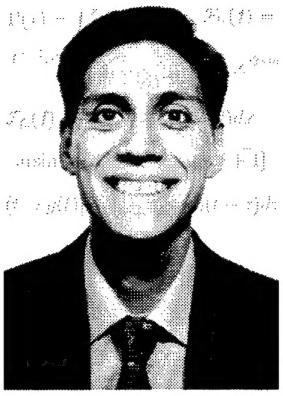
Two assumptions are made in deriving the main results; (i) that the predictive system is linear and (ii) that the noise is uncorrelated. We compare and contrast four different techniques that develop these transfer functions; the vector, matrix, corrected-vector, and surrogate-matrix techniques. The construction of these transfer functions has two main difficulties, measurement noise and measurement collinearity. We find that although the vector technique is easily implemented, it lacks accuracy for ill-conditioned and noisy systems. We show when and why these inaccuracies arise and how to correct them. The matrix technique produces good results, but requires the measurement of external loads during calibration. If a good estimation of noise is available then the corrected-vector technique improves upon the results of the vector technique (however, it still does not match the accuracy achieved by the matrix technique). The surrogate-matrix technique substitutes measurements from additional-sensors for external-loads, achieving an accuracy that matches the matrix technique. Simulations on a simple truss support the theory presented.

The results developed in this report are not merely restricted to helicopters (or indeed aircraft). The results are applicable to any linear system where a predictive capability is to be developed from system measurements.

The investigation herein of noise and redundancy will allow the development of more accurate and more robust transfer functions, and hence rotor loading predictions. In

turn, these accurate loading predictions have the potential to increase safety and reduce operating costs for the Australian Defence Force and other aircraft operators.

Author



Frank G. Polanco
Airframes and Engines Division

Frank Polanco graduated in 1992 with a Bachelor of Aerospace Engineering (Honours) and a Bachelor of Applied Science (Distinction) from the Royal Melbourne Institute of Technology (RMIT). He joined the Aeronautical and Maritime Research Laboratory (AMRL) in 1993, working on aircraft structural integrity and fatigue life monitoring before returning to RMIT to complete a Doctorate in Mathematics. He then rejoined the AMRL in 1998 to work in the area of helicopter life assessment.

Contents

Appendices

Figures

1 Introduction

Accurate prediction of the loading in critical helicopter components has the potential to significantly increase safety and reduce operating costs for the Australian Defence Force and other aircraft operators.

Aircraft component retirement times (CRTs) are calculated using an assumed load spectrum and experimentally-derived fatigue strength data. It is unusual for aircraft to experience loading conditions as severe as those assumed, so the CRTs are normally quite conservative. Nevertheless, Schaefer [28] has found that a few components may require that their CRTs be reduced by up to 75% (for further details see [24]). This possible requirement for CRT reductions has safety implications. On the other hand, due to the conservative nature of the CRT estimations most components are discarded well before their safe fatigue lives have been expended. Hence, the accurate prediction of component loading has two main benefits—increased safety and reduced operational costs.

Most fatigue-critical components in helicopters are located in the rotor system and its drive train [20, p. 46] where, unlike fixed-wing aircraft, they have single load paths and experience high and low cycle fatigue loading [17]. Unfortunately, strain measuring devices are short lived due to the harsh working environment of these critical components. In addition, the transmission of this loading information from rotating components to a recording system is itself a difficult task. In the past this information has been transmitted either via sliprings (resulting in excessive noise problems) or telemetry (normally involving large cumbersome devices). The difficulty in obtaining reliable rotor component loads is demonstrated by the fact that, historically, these loads have only ever been obtained during short-duration experimental trials.

In an earlier report [25] we developed stress transfer functions (STFs), which related the stress in one section of an idealised helicopter truss to the stress in a different section of the helicopter. Once a numerical simulation of this idealised structure was implemented we quickly realised that the strain measurements were sensitive to several variables including strain gauge location and noise in the strain measurements. In fact, under some configurations of strain gauge locations, even a small amount of strain measurement noise completely corrupted the development of the STFs. We concluded the discussion in that report with several open-ended questions re-iterated below. (In the questions shown below we refer to the STF system as "the system", and the *vector* and *matrix* techniques refer to the two procedures developed in an earlier report [25] to determine the STFs.)

- How does noise affect the system's condition number?
- How do structural characteristics affect ill-conditioning?
- Why is the error independent of the approximation order for the vector technique?
- Why did the vector technique under-estimate the results?
- Why does the matrix technique appear to be more stable than the vector technique?
- Are there situations under which the vector technique gives results superior to the matrix technique?
- Will a data analysis method based on the bootstrap method improve results?

In this report we answer these questions, and investigate what effect measurement noise has on predictions of general linear systems. As such, this work is not restricted to merely helicopters (or indeed even aircraft). The results may be applied in any situation where a *predictive* capability is to be developed from measured information.

We begin, in § 2, by revisiting results relating to singular value decomposition. In turn, we use these results to determine the condition number of a two-by-two matrix (which is a measure of "goodness" for gauge configuration on helicopter trusses). We conclude the first section by determining when these gauge configurations are both well- and ill-conditioned.

In the next section (§ 3) we develop a simple truss to investigate problems that may arise in the construction of transfer functions. This simple truss will be used in numerical simulations, where external loads will be applied to the truss and simulated strain gauges will determine the loading in various components. We review the vector and matrix techniques, and also simulation results from earlier work.

The largest section of this report (§ 4) is devoted to the analysis of the least squares (LS) technique in determining the transfer functions, since the LS procedure is the basis of the vector technique. We begin § 4 by investigating the effects of measurement noise on our LS prediction of the transfer function. Two assumptions are made in deriving the main results; (i) that the predictive system is linear and (ii) that the noise is uncorrelated. The first assumption doesn't restrict our analysis to linear systems; we assume only that the predictive system is linear in the measurement variables. For example, the two-dimensional predictive system $h(f,g) = a\,f(x,y) + b\,g(x,y)$ (where a and b are constants) is linear in f and g, even if the functions $f(x,y)$ and $g(x,y)$ are highly non-linear. The second assumption is also non-restrictive since we would assume that any true noise parameter would be independent of noise in all other parameters.

Making these simplifying assumptions we determine when the LS prediction breaks down. A method for correcting errors in LS predictions is then developed, and a sensitivity analysis of this correction undertaken. The results from the preceding subsections are then generalised to higher dimensions. We are then in a position to show why the matrix technique outperforms the vector technique. The relationship between the condition number (of gauge configuration) and the correlation (of the output gauges) is then investigated. Simulation results for the corrected-vector technique and surrogate-matrix technique (two techniques developed to improve results) support the theoretical findings.

The total LS technique is reviewed in § 5. We show the relationship between the LS, corrected LS, and total LS techniques, and determine which of these techniques is best suited to the development of the transfer functions.

2 Condition Number of a 2×2 Matrix

In this section, we first revisit the singular value decomposition technique. The reason for doing so is to determine a measure of "gauge configuration goodness" for our helicopter truss. One such measure is given by the condition number of the gauge configuration on the truss. As such, we determine under what conditions a simple two-by-two matrix becomes both well- and ill-conditioned, where this matrix represents the linear relationship between loads in fixed and dynamic helicopter components. The results from this section will be used to investigate the relationship between system condition number and gauge output correlation.

2.1 Revisiting Singular Value Decomposition

Following the outline in Golub and van Loan [10, p. 71], the singular value decomposition (SVD) of a real m-by-n matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \tag{2.1}$$

where the matrix $\boldsymbol{D}$ is diagonal and the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal, and have the following structure

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m] \in \mathbb{R}^{m\times m}, \qquad \boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n] \in \mathbb{R}^{n\times n},$$

and

$$\boldsymbol{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n},$$

where $p = \min(m, n)$. The concept of a diagonal matrix (that is, a matrix with non-zero terms only along the diagonal) has been extended to rectangular matrices in the above definition of $\boldsymbol{D}$.

The values σ_i for $i = 1, \ldots, p$ are termed singular values. The vectors $\boldsymbol{u}_i$ for $i = 1, \ldots, m$ and $\boldsymbol{v}_i$ for $i = 1, \ldots, n$ are termed left and right singular vectors respectively. For the special case of $\boldsymbol{A}$ real and symmetric, the singular values and absolute value of the eigenvalues of $\boldsymbol{A}$ become identical, as do the singular vectors and eigenvectors. Horn and Johnson [11, p. 213] state that perhaps this is the reason the term *generalised eigenvalue* is occasionally used instead of singular value.

The singular values are ordered so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. The $\operatorname{span}(\boldsymbol{u_1}, \ldots, \boldsymbol{u_r})$ and $\operatorname{span}(\boldsymbol{v_{r+1}}, \ldots, \boldsymbol{v_n})$ define the range and null space of $\boldsymbol{A}$ respectively, where r is the rank of $\boldsymbol{A}$ such that $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$. Thus if $\boldsymbol{A}$ has full rank, then all the singular values are non-zero and $r = p$.

The 2-norm condition number of the matrix $\boldsymbol{A}$, defined in terms of the singular values, is given by

$$\kappa = \frac{\sigma_1}{\sigma_p},$$

that is, the ratio of the largest to smallest singular values. The 2-norm of the matrix $\boldsymbol{A}$ is simply given by the largest singular value, that is $||\boldsymbol{A}||_2 = \sigma_1$. The square of the Frobenius norm is given by the sum of the singular values squared, that is $||\boldsymbol{A}||_F^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2$.

Figure 2.1 shows a geometric interpretation of singular values and vectors for the two-dimensional case. We now see that SVD merely decomposes a matrix into a ro-

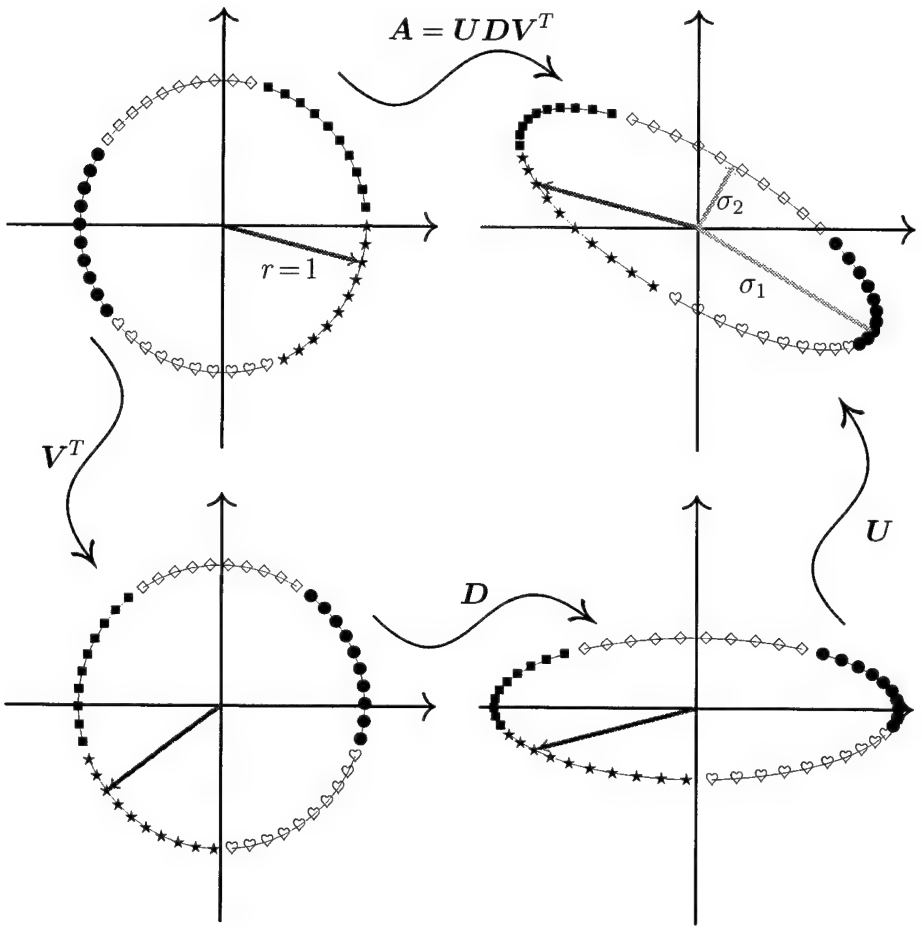


Figure 2.1: Transformation of the unit circle under the mapping of a two-by-two matrix $\boldsymbol{A}$*. The lengths of the semimajor and semiminor axes of the ellipse give the of the maximum and minimum singular values (*σ_1 *and* σ_2*) respectively.*

tation/reflection (the matrix $\boldsymbol{V}^T$), a stretching (the diagonal matrix $\boldsymbol{D}$), and a rotation/reflection (the matrix $\boldsymbol{U}$). The top left figure shows the unit circle, which represents all vectors of unit length; while the top right figure shows the mapping of the unit circle under the matrix $\boldsymbol{A}$ (a linear transformation). The singular values control the length of the ellipse's axes. In general a singular value decomposition may be thought of as a linear mapping from a hyper-sphere to a hyper-ellipsoid, the singular values controlling the lengths of the hyper-ellipsoid's semi-axes. The left and right singular vectors define the shape-preserving transformations. The different symbols drawn onto the circles and ellipses show the effects of the singular vectors (that is, rotations and reflections). The transformation $\boldsymbol{A}$ shown in Figure 2.1 is given by the two-by-two matrix and corresponding

SVD (shown to two decimal places)

$$\boldsymbol{A} = \begin{bmatrix} -1.22 & -0.23 \\ 0.48 & 0.67 \end{bmatrix} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \begin{bmatrix} 0.87 & 0.50 \\ -0.50 & 0.87 \end{bmatrix} \begin{bmatrix} 1.40 & 0.00 \\ 0.00 & 0.50 \end{bmatrix} \begin{bmatrix} -0.92 & -0.38 \\ -0.38 & 0.92 \end{bmatrix}.$$

If r is the rank of $\boldsymbol{A}$ then the least squares (LS) solution of the linear problem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is given by

$$\boldsymbol{x}_{\text{LS}} = \sum_{i=1}^{r} \frac{\boldsymbol{u}_i^T \boldsymbol{b}}{\sigma_i} \boldsymbol{v}_i = \boldsymbol{A}^+ \boldsymbol{b}, \tag{2.2}$$

where $\boldsymbol{A}^+$ is the pseudo-inverse (defined below). The solution $\boldsymbol{x}_{\text{LS}}$ minimises the 2-norm residual $||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2$, and the minimum is given by

$$||\boldsymbol{A}\boldsymbol{x}_{\text{LS}} - \boldsymbol{b}||_2^2 = \sum_{i=r+1}^{m} \left(\boldsymbol{u}_i^T \boldsymbol{b}\right)^2.$$

The pseudo-inverse is defined as

$$\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{D}^+\boldsymbol{U}^T,$$

where

$$\boldsymbol{D}^+ = \text{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right) \in \mathbb{R}^{n \times m}.$$

If the rank$(\boldsymbol{A}) = n$, then $\boldsymbol{A}^+ = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$, while if $m = n = \text{rank}(\boldsymbol{A})$, then $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$. Typically $\boldsymbol{A}^+$ is defined to be the unique matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ that satisfies the four Moore-Penrose conditions (see Golub and van Loan [10, p. 243]). The four conditions amount to the requirement that $\boldsymbol{A}\boldsymbol{A}^+$ and $\boldsymbol{A}^+\boldsymbol{A}$ be orthogonal projections onto the range$(\boldsymbol{A})$ and range$(\boldsymbol{A}^T)$ respectively. Schmidt [29] terms the matrix $\boldsymbol{A}^+$ that satisfies the Moore-Penrose conditions the generalised inverse. Golub and van Loan [10] use the pseudo-inverse to show that small changes in $\boldsymbol{A}$ or $\boldsymbol{b}$ can induce arbitrarily large changes in the least squares solution.

One approach to determine the numerical rank of $\boldsymbol{A}$ is to assign a tolerance δ such that the singular values are split into two sets

$$\sigma_1 \geq \cdots \geq \sigma_r > \delta \geq \sigma_{r+1} \geq \cdots \geq \sigma_n,$$

and if the matrix $\boldsymbol{A}$ elements have n_{sd} significant digits one suggested choice for the tolerance is $\delta = 10^{-n_{sd}}||\boldsymbol{A}||_\infty$. The infinity norm of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ with elements a_{ij} is defined as

$$||\boldsymbol{A}||_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|.$$

We have seen that the SVD decomposition of a matrix is simply an orthogonalisation of the matrix into a sum of rank one matrices (see Equation (2.2)). From this decomposition we gain such useful information as the matrix's rank, distance to the nearest singular matrix, condition number, several matrix norms, LS solution, error in the LS solution, pseudo-inverse, and the directions of the range and null space.

2.2 Using SVD to Determine the Condition Number

In this subsection we determine the condition number of a two-by-two matrix using the SVD results from the previous subsection.

From the previous subsection on SVD we know that every matrix $\boldsymbol{A}$ has a SVD given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal and $\boldsymbol{D}$ is diagonal.

Let the two-by-two matrix $\boldsymbol{A}$ have the form

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{2.3}$$

One way to express the SVD of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}, \tag{2.4}$$

since the matrices involving sinusoidal functions are orthogonal, and obviously the central matrix is diagonal. Multiplying out the above expression, making the change of notation substitutions $\widehat{\sigma}_1 = \sigma_1 \cos\theta \cos\alpha$, $\widehat{\sigma}_2 = \sigma_2 \cos\theta \cos\alpha$, $x = \tan\theta$, and $y = \tan\alpha$, and equating the result to $\boldsymbol{A}$ yields the four equations

$$\widehat{\sigma}_1 - \widehat{\sigma}_2 x y = a, \tag{2.5}$$

$$\widehat{\sigma}_1 y + \widehat{\sigma}_2 x = b, \tag{2.6}$$

$$-\widehat{\sigma}_1 x - \widehat{\sigma}_2 y = c, \qquad \text{and} \tag{2.7}$$

$$-\widehat{\sigma}_1 x y + \widehat{\sigma}_2 = d. \tag{2.8}$$

Solving for $\widehat{\sigma}_1$ using Equations (2.5) and (2.7) gives

$$\widehat{\sigma}_1 = \frac{a - cx}{1 + x^2}. \tag{2.9}$$

Similarly using Equations (2.6) and (2.8) to solve for $\widehat{\sigma}_2$ gives

$$\widehat{\sigma}_2 = \frac{d + bx}{1 + x^2}. \tag{2.10}$$

Substituting the two equations shown above into Equations (2.6) and (2.7), and solving for y gives

$$y = \frac{b - dx}{a - cx} \qquad \text{and} \qquad y = -\frac{c + ax}{d + bx}$$

respectively. Eliminating y using the above two expressions and solving for x yields the two solutions

$$x_+ = z + \sqrt{1 + z^2} \qquad \text{and} \qquad x_- = z - \sqrt{1 + z^2}, \tag{2.11}$$

where

$$z = \frac{a^2 + b^2 - c^2 - d^2}{2(ac + bd)}. \tag{2.12}$$

Now we want to determine the condition number κ, which is merely the ratio of the largest to smallest singular values of the matrix. We do not know, at this stage, which

of $\widehat{\sigma}_1$ or $\widehat{\sigma}_2$ is the larger. Let's denote the first possibility as $\kappa_{12}(x) = \sigma_1/\sigma_2$ and the second possibility as $\kappa_{21}(x) = \sigma_2/\sigma_1$. (Note that from the definitions of $\widehat{\sigma}$ we know that $\widehat{\sigma}_1/\widehat{\sigma}_2 = \sigma_1/\sigma_2$.) It can be shown that $\kappa_{12}(x_-) = \kappa_{21}(x_+)$ and that $\kappa_{12}(x_+) = \kappa_{21}(x_-)$ (which incidentally explains why we have two solutions for x), and hence from now on we only consider κ_{12}.

We can write down κ_{12} using Equations (2.9) and (2.10) as

$$\kappa_{12}(x) = \frac{a - cx}{d + bx},$$

and then substituting Equations (2.11) and (2.12) gives

$$\kappa_{12} = \frac{a^2 + b^2 + c^2 + d^2 \pm \sqrt{\left[(a-d)^2 + (b+c)^2\right]\left[(a+d)^2 + (b-c)^2\right]}}{2(ad - bc)}.$$

It is the positive sign in front of the radical that maximises the above expression, and hence we take the positive case. Also note that the sign of the above expression depends on the sign of the determinant of $\boldsymbol{A}$ (that is, the sign of $ad - bc$). Remember that the formulation initially involved two frequency parameters θ and α (see Equation (2.4)). Adding π to either θ or α yields the same magnitude singular values, but with opposite sign. However, we are only interested in the magnitude and hence take the absolute value of the denominator in the final expression of the condition number shown below

$$\kappa = \frac{a^2 + b^2 + c^2 + d^2 + \sqrt{\left[(a-d)^2 + (b+c)^2\right]\left[(a+d)^2 + (b-c)^2\right]}}{2|ad - bc|}.$$

The condition number developed in this section suffers from complexity, especially when we try to determine under what conditions the matrix is well-conditioned. As such, a simpler formulation is attempted below.

2.3 Using Norms to Determine the Condition Number

In this subsection we develop the condition number of a two-by-two matrix $\boldsymbol{A}$, with the form shown in Equation (2.3), using the 2-norm. This formulation has the advantage of simplicity, but we loose some information, namely the null space directions given by the angle α in the previous section. However, since we are only interested in the condition number, this loss of information is of no concern. If necessary the angle α can be obtained by solving the simple system $\boldsymbol{V}^T = (\boldsymbol{UD})^{-1}\boldsymbol{A}$.

We know that the condition number may be expressed as the ratio of singular values, namely $\kappa = \sigma_{\max}/\sigma_{\min}$. But we also know that the maximum and minimum singular values are related to the 2-norm (see Golub and van Loan [10] for example) by

$$\sigma_{\max} = \max_{||\boldsymbol{x}||_2=1} ||\boldsymbol{Ax}||_2 \qquad \text{and} \qquad \sigma_{\min} = \min_{||\boldsymbol{x}||_2=1} ||\boldsymbol{Ax}||_2. \tag{2.13}$$

Remember that for the two-dimensional case these singular values have geometric interpretations. We are either maximising or minimising over the unit circle, $||\boldsymbol{x}||_2 = 1$. The

matrix $\boldsymbol{A}$ maps the unit circle onto an ellipse, which has semimajor and semiminor axes given by the maximum and minimum singular values (see Figure 2.1).

Now, if the vector $\boldsymbol{x}$ is of unit length, then the transformation of $\boldsymbol{x}$ under $\boldsymbol{A}$ may be written as

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}.$$

Remember that the 2-norm (or Euclidean norm) of a vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$ is given by $||\boldsymbol{x}||_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$. Let ϵ be the square of the 2-norm of the vector $\boldsymbol{A}\boldsymbol{x}$, that is $\epsilon = ||\boldsymbol{A}\boldsymbol{x}||_2^2$. Then using the trigonometric relations $\cos^2\theta = (1 + \cos 2\theta)/2$ and $\sin\theta\cos\theta = (\sin 2\theta)/2$, after a small amount of simplification we can write down ϵ as

$$\epsilon = \tfrac{1}{2}(a^2 + b^2 + c^2 + d^2) + \tfrac{1}{2}(a^2 - b^2 + c^2 - d^2)\cos 2\theta + (ab + cd)\sin 2\theta. \tag{2.14}$$

We want to both maximise and minimise ϵ (see Equation (2.13)). Thus we differentiate ϵ with respect to θ, and then solve for $d\epsilon/d\theta = 0$ giving

$$\tan 2\theta = \frac{\gamma_1}{\gamma_2},$$

where

$$\begin{aligned} \gamma_1 &= 2(ab + cd), \\ \gamma_2 &= a^2 - b^2 + c^2 - d^2, \quad \text{and} \\ \gamma_3 &= a^2 + b^2 + c^2 + d^2. \end{aligned} \tag{2.15}$$

(We will use the definition for γ_3 later.) Using trigonometry, if $\tan 2\theta = \gamma_1/\gamma_2$ then

$$\cos 2\theta = \frac{\gamma_2}{\sqrt{\gamma_1^2 + \gamma_2^2}} \qquad \text{and} \qquad \sin 2\theta = \frac{\gamma_1}{\sqrt{\gamma_1^2 + \gamma_2^2}}. \tag{2.16}$$

Substituting these two expressions into Equation (2.14) and simplifying yields

$$\epsilon_1 = \frac{1}{2}\left(\gamma_3 + \sqrt{\gamma_1^2 + \gamma_2^2}\right),$$

which is the square of either the maximum or minimum singular value. We will show later that ϵ_1 is in fact the square of the maximum singular value. Thus to find the minimum all we need do is add $\pi/2$ to the value of θ (see Figure 2.1). From trigonometry we know that $\sin[2(\theta + \pi/2)] = -\sin 2\theta$ and $\cos[2(\theta + \pi/2)] = -\cos 2\theta$. Thus substituting the negative forms of Equation (2.16) into ϵ, given by Equation (2.14), yields

$$\epsilon_2 = \frac{1}{2}\left(\gamma_3 - \sqrt{\gamma_1^2 + \gamma_2^2}\right).$$

Since γ_3 is always positive, whenever $\boldsymbol{A} \neq \boldsymbol{0}$, we now see that $\epsilon_1 \geq \epsilon_2$, and hence $\sigma_{\max} = \sqrt{\epsilon_1}$ and $\sigma_{\min} = \sqrt{\epsilon_2}$. The square of the condition number is then simply

$$\begin{aligned} \kappa^2 &= \frac{\epsilon_1}{\epsilon_2} \\ &= 1 + \frac{2}{\dfrac{\gamma_3}{\sqrt{\gamma_1^2 + \gamma_2^2}} - 1}. \end{aligned} \tag{2.17}$$

Finally, the condition number in terms of the matrix elements is given by

$$\kappa = \sqrt{1 + \frac{2}{\frac{a^2+b^2+c^2+d^2}{\sqrt{4(ab+cd)^2+(a^2-b^2+c^2-d^2)^2}} - 1}}. \tag{2.18}$$

It is interesting to note that the γ_i may be written compactly in terms of the columns of matrix $\boldsymbol{A}$. Let $\boldsymbol{u} = [a, c]^T$ and $\boldsymbol{v} = [b, d]^T$ be the first and second columns of $\boldsymbol{A}$ respectively, that is $\boldsymbol{A} = [\boldsymbol{u}, \boldsymbol{v}]$, then $\gamma_1 = 2\boldsymbol{u}^T\boldsymbol{v}$, $\gamma_2 = \boldsymbol{u}^T\boldsymbol{u} - \boldsymbol{v}^T\boldsymbol{v}$, and $\gamma_3 = \boldsymbol{u}^T\boldsymbol{u} + \boldsymbol{v}^T\boldsymbol{v}$.

We now know that the condition number of a two-by-two matrix is given by Equation (2.18), which allows us to investigate the relation between gauge configuration and ill-conditioning.

2.4 Special Cases: Perfect- and Ill-Conditioning

We want to determine when the matrix $\boldsymbol{A}$ is perfectly-conditioned ($\kappa = 1$, which incidentally is the minimum), and when it is ill-conditioned ($\kappa \to \infty$).

We will consider the ill-conditioned case first since it is straight forward. From Equation (2.17) we see that $\kappa \to \infty$ as $\gamma_3^2/(\gamma_1^2 + \gamma_2^2) \to 1$. Simplifying the latter expression we end up with the condition $ad = bc$, which is satisfied whenever the determinant is zero, that is $|\boldsymbol{A}| = 0$.

We now consider the perfectly conditioned case. In order for κ to be unity we require that the expression $\gamma_3/\sqrt{\gamma_1^2 + \gamma_2^2}$ tend to infinity. Note that this expression is strictly non-negative, and hence we need not consider the case of negative infinity. For non-zero matrices ($\boldsymbol{A} \neq \boldsymbol{0}$) we know that γ_3 is positive (see Equation (2.15)), and hence $\kappa = 1$ implies that $\gamma_1^2 + \gamma_2^2 = 0$. Solving for the matrix element a gives the four solutions

$$a = -d - \boldsymbol{i}|b - c|, \qquad a = -d + \boldsymbol{i}|b - c|, \qquad a = d - \boldsymbol{i}|b + c|, \quad \text{and} \quad a = d + \boldsymbol{i}|b + c|,$$

where $\boldsymbol{i} = \sqrt{-1}$. Similar expressions are obtained if we solve for one of the other elements b, c, or d instead of solving for a. Notice that all these solutions have an imaginary component, unless $b = \pm c$. Thus if all the matrix elements are real, then the only way to achieve perfect conditioning is if $b = \pm c$ and $a = \mp d$. That is, $\kappa = 1$ only when the matrix $\boldsymbol{A}$ has the form

$$\begin{bmatrix} a & b \\ \pm b & \mp a \end{bmatrix}. \tag{2.19}$$

We found that a two-by-two matrix is ill-conditioned when the determinant is zero, and perfectly-conditioned when it has the form in Equation (2.19). (Note, however, that a small determinant does not imply an ill-conditioned system; and reflexively, a large determinant does not imply a well-conditioned system [10, p. 81] [33, p. 70].)

3 Solution of a Determinate Simple Truss

In this section we numerically simulate the stress in a simple truss. The rationale for this simulation is to demonstrate that, in a real helicopter, it is possible to predict the stress in a rotor component from strain gauge measurements on fixed airframe members. The simple simulation developed in this section only looks at effects we are currently interested in and ignores effects we are not yet modelling (for example, vibration).

Although we only investigate the effects of rotor loads on the stress, we can easily take non-load related measurements, such as airspeed and accelerations. In fact, provided we know the functional form between the stress at the rotor component and this measured parameter, we can still develop a linear model, which is exactly what we do in this section.

3.1 Comparison of Vector and Matrix Techniques

We noted some anomalous behaviour by a ten member truss in an earlier report [25]. Namely, two apparently similar inversion techniques produced different results when the underlying system of equations took on certain ill-conditioned forms (in particular collinearity).

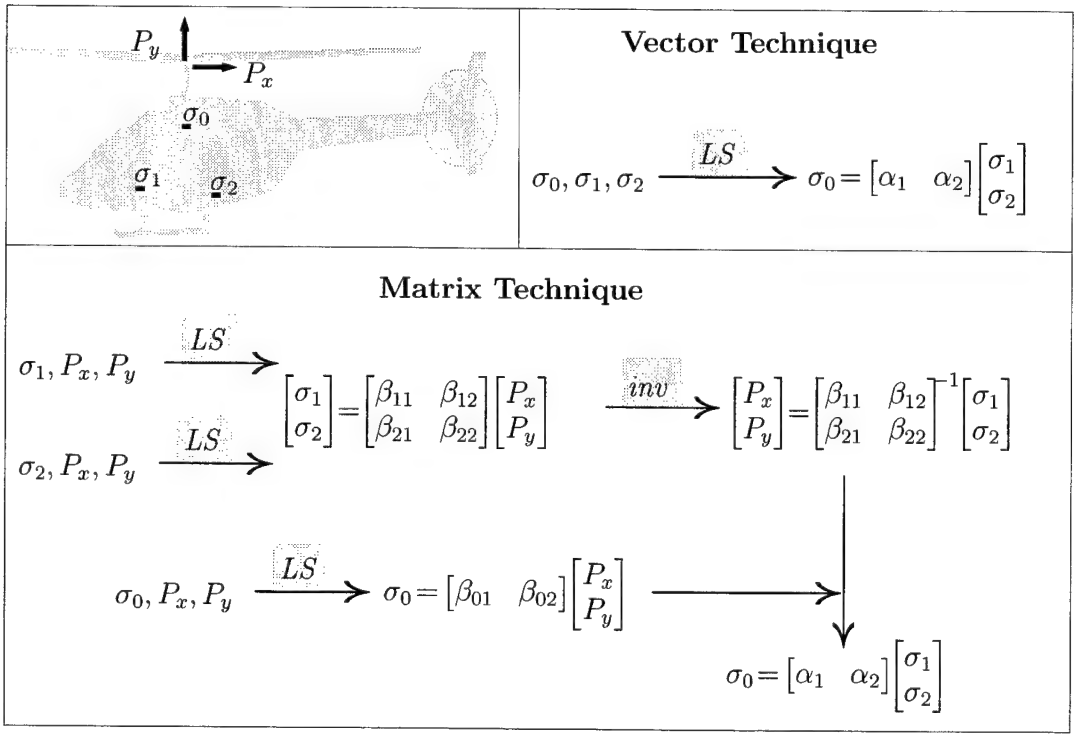


Figure 3.1: Schematic of the vector and matrix techniques. P_x and P_y are external loads and the σ_i are stresses. The operators LS *and* inv *represent the least squares solution and inversion procedure respectively.*

Figure 3.1 schematically illustrates the main procedures in both these techniques, termed the *vector* and *matrix* techniques. Note that the final form of the solution is the same for both techniques, namely a linear model of the stress at the zeroth gauge (σ_0) using the stresses at the first and second strain gauges (σ_1 and σ_2 respectively). However, the matrix technique additionally uses measured external loading (P_x and P_y) as an intermediate step in determining the final solution (that is, during calibration). Once the constant coefficients α_1 and α_2 have been determined, the external loading (represented by P_x and P_y) is no longer required.

The coefficients α_i in the vector and matrix techniques, and β_{ij} in the matrix technique are all constant and determined from the least squares (LS) solution. The operators (depicted above the arrows in shaded boxes) are abbreviated by *LS* for the least squares solution and *inv* for the inversion procedure. For the simple example depicted in this schema we know that α_1 and α_2 for the matrix technique are given respectively by

$$\alpha_1 = \frac{\beta_{01}\beta_{22} - \beta_{02}\beta_{21}}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}} \tag{3.1}$$

and

$$\alpha_2 = \frac{\beta_{02}\beta_{11} - \beta_{01}\beta_{12}}{\beta_{11}\beta_{22} - \beta_{12}\beta_{21}}. \tag{3.2}$$

The vector technique involves determining the LS fit to the m data points $(\sigma_{0i}, \sigma_{1i}, \sigma_{2i})$ for $i = 1, 2, \ldots, m$. The matrix technique involves determining LS fits of all three stresses (σ_0, σ_1, and σ_2) using the external loads (P_x and P_y). Thus, given the data set $(P_{xi}, P_{yi}, \sigma_{0i}, \sigma_{1i}, \sigma_{2i})$ for $i = 1, 2, \ldots, m$, we can determine the LS fits for the three stresses. We first invert the system of equations involving the stresses σ_1 and σ_2, so that the external loadings (P_x and P_y) are now given as a function of σ_1 and σ_2. Substituting these expressions for P_x and P_y into the LS solution for σ_0 in terms of P_x and P_y yields the desired transfer function relating the stresses σ_1 and σ_2 to the stress σ_0 (see Figure 3.1).

An alternative, geometric, interpretation of the vector and matrix techniques is depicted in Figure 3.2. As can be seen for the vector technique, Figure 3.2(a), the plane that predicts the stress at the zeroth gauge is determined using LS. (The LS technique merely minimises the sum of distances squared between the data points and the resulting plane.) Points above the plane are shown in a light grey, while points below the plane are depicted in a darker grey. Notice that the plane (by its very nature) linearly relates the stress in the zeroth gauge to the stresses in the first and second gauges, that is $\sigma_0 = \alpha_1\sigma_1 + \alpha_2\sigma_2$. Once the LS σ_0 plane is determined, then for any given stresses $\sigma_1 = \sigma_1^\star$ and $\sigma_2 = \sigma_2^\star$ the stress at the zeroth gauge is given by $\sigma_0^\star$, the height to the LS σ_0 plane (see Figure 3.2).

In contrast, the matrix technique predicts the stress at the zeroth gauge using a more convoluted process. For clarity, the data points used to construct the LS planes are omitted in Figure 3.2(b) (the matrix technique). Three LS planes are determined using the two external forces P_x and P_y and the three stresses σ_0, σ_1, and σ_2. Now, given the stress $\sigma_1 = \sigma_1^\star$ we determine the line (in the P_x-P_y plane) that causes the LS σ_1 plane to intersect the horizontal plane $\sigma_1^\star$. Similarly, given the stress $\sigma_2 = \sigma_2^\star$ we determine the line in the P_x-P_y plane that causes the LS σ_2 plane to intersect the horizontal plane $\sigma_2^\star$. The intersection of these two lines yields a coordinate $(P_x^\star, P_y^\star)$, which is used to evaluate

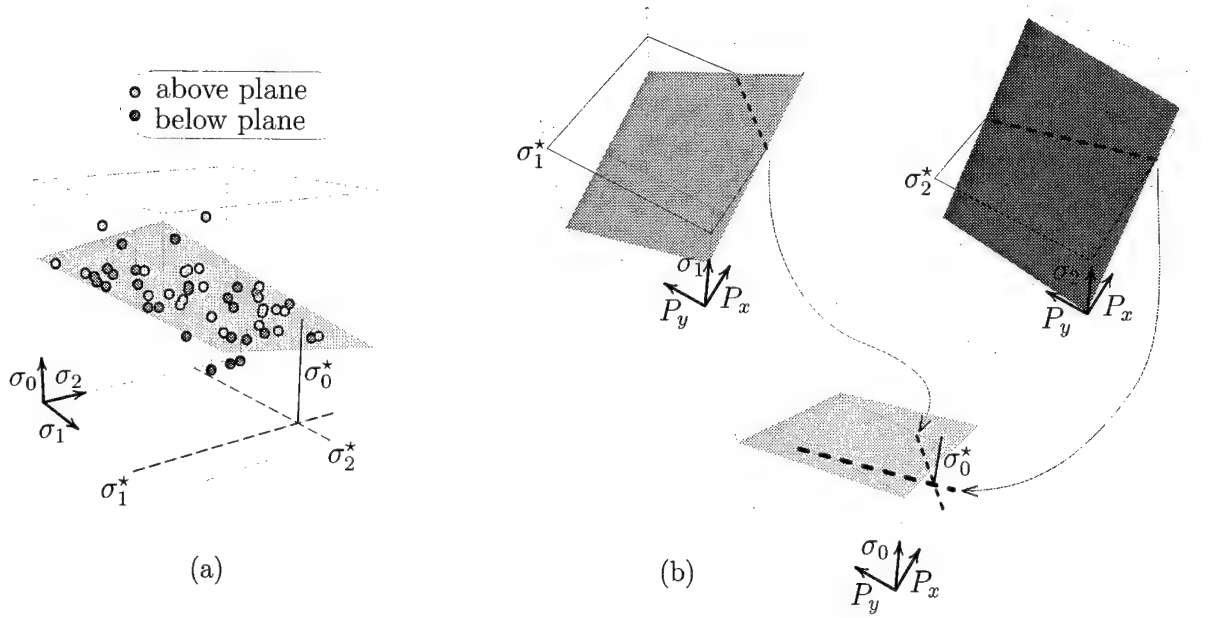


Figure 3.2: Geometric interpretation of the (a) vector and (b) matrix techniques. For clarity the points that define the LS planes are omitted from the matrix technique illustration (b).

the stress at the zeroth gauge $\sigma_0^\star$ using the LS σ_0 plane. In summary, given $\sigma_1^\star$ and $\sigma_2^\star$ we determine the coordinate $(P_x^\star, P_y^\star)$. This coordinate allows us to evaluate $\sigma_0^\star$ from the LS σ_0 plane.

We have seen that the final functional form of the vector and matrix techniques is identical, the difference between the methods being the procedure used to obtain that result. In the vector technique we use the LS solution to develop a stress transfer function between strain gauges on the truss; no external loading information is required. The matrix method also uses the LS procedure, but additionally requires external loading information during calibration.

In order to simplify the idealised structure from the earlier report [25], we will now develop a determinate truss with the minimum number of members possible.

3.2 Solution of a Three Member Truss

Consider the three member statically determinate truss shown in Figure 3.3. Joint O is pinned to the ground and joint B is connected to the ground by a roller bearing, all joints are pinned. Joint C has both a horizontal load P_x and a vertical load P_y. The three strain gauges are denoted by G_{OB} on member OB, G_{OC} on member OC, and G_{BC} on member BC.

Denote the horizontal and vertical distance from the origin (at joint O) to joint C by x_C and y_C respectively. Similarly denote the horizontal distance from joint O to joint B by x_B. Let's non-dimensionalise these distances by the vertical distance y_C, yielding the

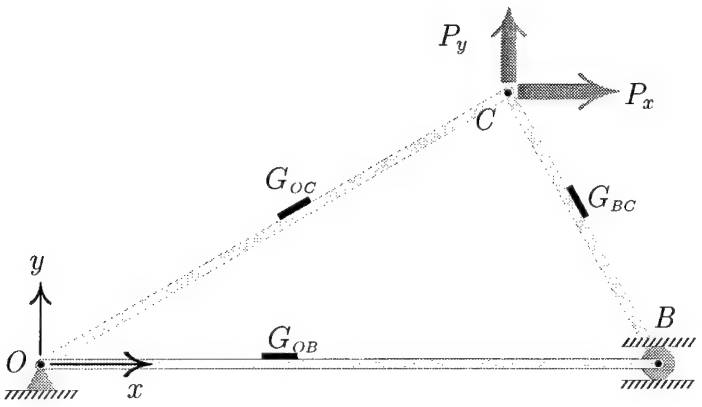


Figure 3.3: Three member statically determinate truss, with external loading P_x and P_y. The three strain gauges are denoted by G_{OB}, G_{OC}, and G_{BC}.

non-dimensional quantities

$$\beta = \frac{x_B}{y_C} \qquad \text{and} \qquad \gamma = \frac{x_C}{y_C}.$$

We can now express any enclosed angle within this triangle using non-dimensional quantities β and γ. So, for example, the cosine of the angle enclosing the joints O, B, and C is given by $\cos\theta_{OBC} = (\beta-\gamma)/\sqrt{1+(\beta-\gamma)^2}$. Furthermore, let all three members have unit cross-sectional area.

Solving for the forces in each member, then dividing by the cross-sectional area (unity in our case) gives us the three stresses

$$\sigma_{OB} = (1-\gamma/\beta)(P_x - \gamma P_y),$$
$$\sigma_{OC} = \frac{\sqrt{1+\gamma^2}}{\beta}\left[P_x + (\beta-\gamma)P_y\right],$$

and

$$\sigma_{BC} = -\frac{\sqrt{1+(\beta-\gamma)^2}}{\beta}\left(P_x - \gamma P_y\right).$$

We see that the stresses at gauges G_{OB} and G_{BC} are just multiples of each other. In fact, a little reflection on the problem will show that there is no combination of two external forces that will generate a "proper" system of equations. Let's keep the same triangular truss shown in Figure 3.3, but shift the external forces to different joints (not necessarily applying P_x and P_y at the same joint).

Applying either external force P_x or P_y at joint O is pointless since the truss will not experience any stress. Similarly, the roller bearing (not the truss) will take out any vertical force applied at joint B. Thus apart from the loading shown in Figure 3.3, there is only one other external loading configuration that will produce what appears to be "sensible" results; namely applying a vertical force at joint C and a horizontal force at

joint B. However, under this loading configuration the stresses at gauges G_{OC} and G_{BC} are only dependent on the vertical loading, and hence independent of the horizontal loading at joint B. Thus the stress at gauge G_{OB} cannot be determined from the stresses at gauges G_{OC} and G_{BC} alone under this loading configuration.

The only way to obtain a sensible result would be to apply either (or both) of the external loads somewhere on the member BC (excluding the end points). This loading configuration would generate a bending moment within member BC, which would be dependent on both where the strain is measured G_{BC} and the locations of the external loads applied on member BC. (To determine stress we would additionally require structural information such as the second moment of area and the distance from the neutral axis to the strain gauge.)

So as not to obfuscate the properties of this problem with unnecessary information, a five member truss is developed in the next section. (Note that a suitable four member statically determinate truss, without at least one of the members undergoing bending, could not be constructed. Hence we construct a five member truss in the next section.)

3.3 Solution of a Five Member Truss

In this section we will investigate the vector and matrix solution techniques using the five member statically determinate truss shown in Figure 3.4. As before, all joints are pinned, and joints O and B are grounded via a pin joint and a roller bearing respectively.

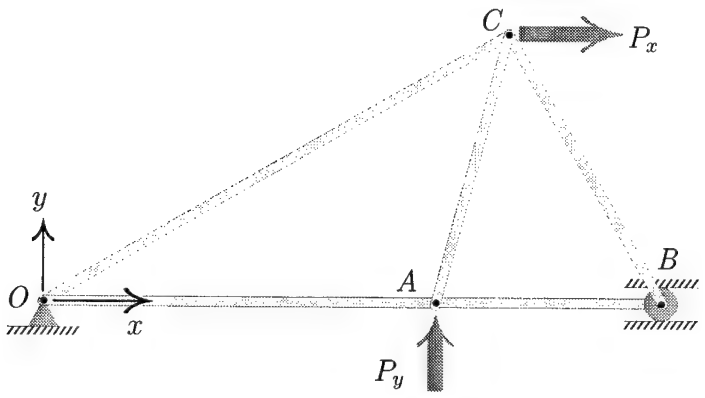


Figure 3.4: Five member statically determinate truss, with external loading P_x and P_y.

Note that if the vertical load P_y was applied at joint C instead of joint A, then member AC would experience no stress, reducing the truss to the three member configuration shown in Figure 3.3. (This fact is easily verified by considering the sum of vertical forces on joint A.)

Again, denote the horizontal and vertical distances from the origin (joint O) to joint C

by x_C and y_C respectively. Similarly, the horizontal distances from the origin to joints A and B are given by x_A and x_B respectively. Non-dimensionalising all distances by the vertical height y_C yields the following non-dimensional quantities

$$\alpha = \frac{x_A}{y_C}, \qquad \beta = \frac{x_B}{y_C}, \quad \text{and} \quad \gamma = \frac{x_C}{y_C}.$$

To maintain the general geometry shown in Figure 3.4 we need to limit the ranges of the lengths x_a, x_b, and y_c. The conditions $y_c \neq 0$ and $0 < |\alpha| < |\beta|$ together with $\alpha\beta > 0$ satisfies this criterion.

Now, using simple trigonometric relations both the sine and cosine of any truss angle may be written in terms of α, β, and γ. For example, the cosine of the angle made by joints ABC is given by $\cos\theta_{ABC} = (\beta - \gamma)/\sqrt{1 + (\beta - \gamma)^2}$. Using these trigonometric relationship, we can develop the loads in every truss member, and (provided we have cross-sectional areas) also the stresses.

If all the truss members have unit cross-sectional area, then the stress in the five members are

$$\sigma_{OA} = (1 - \gamma/\beta)P_x - \gamma(1 - \alpha/\beta)P_y, \tag{3.3}$$

$$\sigma_{OC} = \frac{\sqrt{1+\gamma^2}}{\beta}[P_x + (\beta - \alpha)P_y], \tag{3.4}$$

$$\sigma_{AB} = (1 - \gamma/\beta)(P_x - \alpha P_y), \tag{3.5}$$

$$\sigma_{AC} = -\sqrt{1 + (\gamma - \alpha)^2}\, P_y, \tag{3.6}$$

and

$$\sigma_{BC} = -\frac{\sqrt{1 + (\beta - \gamma)^2}}{\beta}(P_x - \alpha P_y), \tag{3.7}$$

where, for example, σ_{AB} is the stress in member AB.

Like the three member truss, the stress in member AB is a multiple of the stress in member BC (unless $\beta = \gamma$). Other special cases are obtained for different combinations of α, β, and γ. Figure 3.5 shows the five member truss under different geometric configurations. Notice the two special configurations (a) $\alpha = \gamma$ and (b) $\beta = \gamma$. If member AC is perpendicular to member OB (that is, $\alpha = \gamma$), then the stress in members OA and AB are equal, and both are proportional to the stress in member BC (that is, $\sigma_{OA} = \sigma_{AB} \propto \sigma_{BC}$). While if member BC is perpendicular to member OB (that is, $\beta = \gamma$), then the stress in members OA and AC are proportional, and the stress in member AB is zero (that is, $\sigma_{OA} \propto \sigma_{AC}$ and $\sigma_{AB} = 0$).

To obtain sensible stress solutions for a truss with two external loads, the minimum number of beam members found to be suitable was five. For truss beams of varying length (and hence geometries), we developed the stress equations for each member of the truss. We will use this truss in simulations to support theoretical results obtained in the remainder of this report.

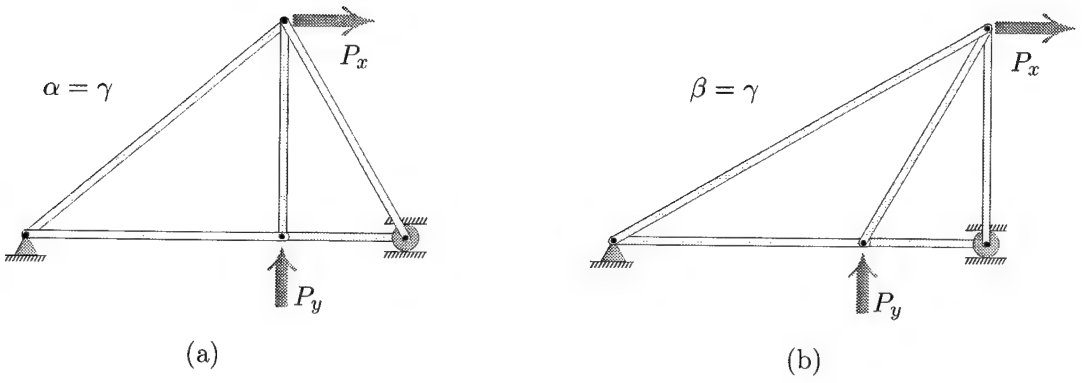


Figure 3.5: Five member truss under two special cases (a) $\alpha = \gamma$ and (b) $\beta = \gamma$.

3.4 Simulation of a Five Member Truss

The results presented in this subsection have already been reported elsewhere [25], and hence we summarise some of the main findings.

Consider the five member truss developed in § 3.3 undergoing random loading, that is, P_x and P_y are random and independent forces. We will simulate the stress in one truss member using the stress at another two truss members, we choose $\sigma_{OC} = f(\sigma_{OA}, \sigma_{AB})$. This combination of members was chosen for simulation through a process of elimination. As shown in the previous section, the stresses in members AB and BC are proportional and hence could not be used concurrently in this sort of simulation. Member AC is unaffected by the horizontal loading P_x, and hence would make this sort of simulation relatively simple. Given this information, the only two choices remaining were (i) choosing member AB or member BC and (ii) choosing which of the three chosen members would be the dependent member. The decisions to choose AB instead of BC, and choose OC as the dependent member were made for the following reason; under the above truss geometry members OA and AB should have a high degree of correlation. This correlation would allow us to investigate the effects of system collinearity and noise on stress estimation.

Using Equations (3.3)–(3.5) we can solve for the stress in member OC in terms of the stresses in members OA and AB, thus giving

$$\begin{aligned}\sigma_{OC} &= \frac{\sqrt{1+\gamma^2}}{\alpha-\gamma}\left[\sigma_{OA} + \left(\frac{\alpha-\beta}{\beta-\gamma}\right)\sigma_{AB}\right] \\ &= a_1\,\sigma_{OA} + a_2\,\sigma_{AB}. \end{aligned} \tag{3.8}$$

In the above expression σ_{OA} and σ_{AB} are the system inputs and σ_{OC} is the system output. If the geometry of the truss is fixed by setting the non-dimensional lengths $\alpha = 1.000$, $\beta = 2.000$, and $\gamma = 1.100$, then the above equation with four digit precision becomes

$$\sigma_{OC} = -14.87\sigma_{OA} + 16.52\sigma_{AB}. \tag{3.9}$$

The above geometry results in a truss with members OA and AB having the same length, and joint C being slightly to the right of joint A.

If both inputs (σ_{OA} and σ_{ab}) contain no noise, then the least squares (LS) method would recover the above expression (provided enough measurement points were used). However, when there is noise in the input measurements, the vector method produces results with increasing error as the number of points in the LS approximation increases. On the contrary, the matrix method produces results with decreasing error as the number of points in the LS approximation increases.

We now describe a numerical simulation of this five-member truss. Using random input loads P_x and P_y in Equations (3.3)–(3.5) we obtain the exact stress in the three members OA, OC, and AB. We add white noise in the range $\pm 10\%$ to these exact loads and exact stresses to obtain our measured results $\widehat{P}_x$, $\widehat{P}_y$, $\widehat{\sigma}_{OA}$, $\widehat{\sigma}_{OC}$, and $\widehat{\sigma}_{AB}$. For example, the measured stress in the member OA is $\widehat{\sigma}_{OA} = \sigma_{OA}(1+\epsilon)$, where $\epsilon \in [-0.10, +0.10]$ is a random number with uniform distribution (white noise).

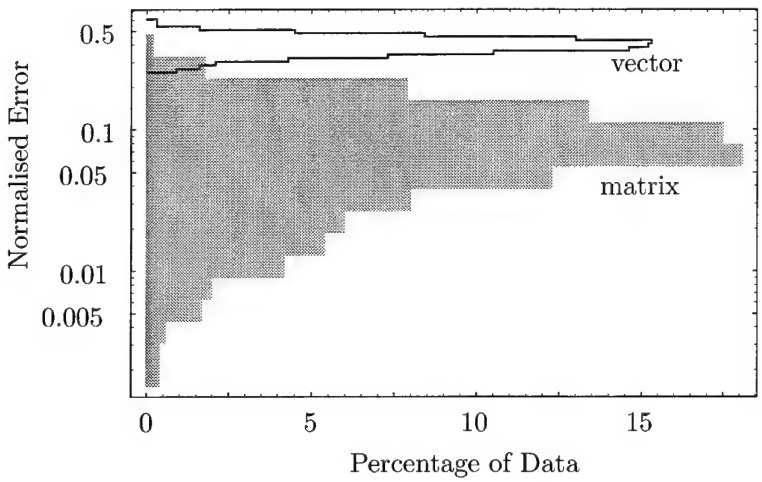


Figure 3.6: Distribution of LS errors for 1000 different measurement sample sets. Each sample set contains 128 data points. (Results from an earlier report [25].)

Figure 3.6 shows the distribution (or frequency plot) of normalised errors when the LS method is used to develop an approximation to Equation (3.9). The vertical scale measures the error of the LS approximation (on a logarithmic scale), while the horizontal scale shows how much of the data falls into each error bin. The distribution was partitioned into 16 data bins of equal width on a logarithmic scale. The distribution includes 1000 measurement samples (or batches); each measurement sample set contains 128 data points. In other words 128 points were used for each LS approximation (the order of the LS approximation), and there were 1000 LS approximations (the number of batches in the distribution). By data *point* we mean an input measurement $(\widehat{P}_x, \widehat{P}_y, \widehat{\sigma}_{OA}, \widehat{\sigma}_{OC}, \widehat{\sigma}_{AB})$ at a particular point in time. The vector technique does not use the external loading information $\widehat{P}_x$ and $\widehat{P}_y$ at all. In contrast the matrix technique only uses this loading information to determine the coefficient in Equation (3.9) (that is, during calibration). Once these coefficients have been determined, the matrix technique no longer requires this external loading information.

We have defined the normalised error of the LS method as

$$\text{normalised error} = \frac{||\widehat{\boldsymbol{a}} - \boldsymbol{a}||_2}{||\boldsymbol{a}||_2},$$

where $\boldsymbol{a} = [a_1, a_2]^T$ and $\widehat{\boldsymbol{a}} = [\widehat{a}_1, \widehat{a}_2]^T$ are the exact and approximate coefficient vectors for Equation (3.8) and $||\boldsymbol{x}||_2^2 = \boldsymbol{x}^T\boldsymbol{x}$ is the 2-norm (or Euclidean norm) of the vector $\boldsymbol{x}$.

As can be seen from Figure 3.6, for a 128 point LS approximation the matrix technique produces more accurate results on average. Both vector and matrix solutions have approximately a log-normal distribution of errors, with a small amount of skewness.

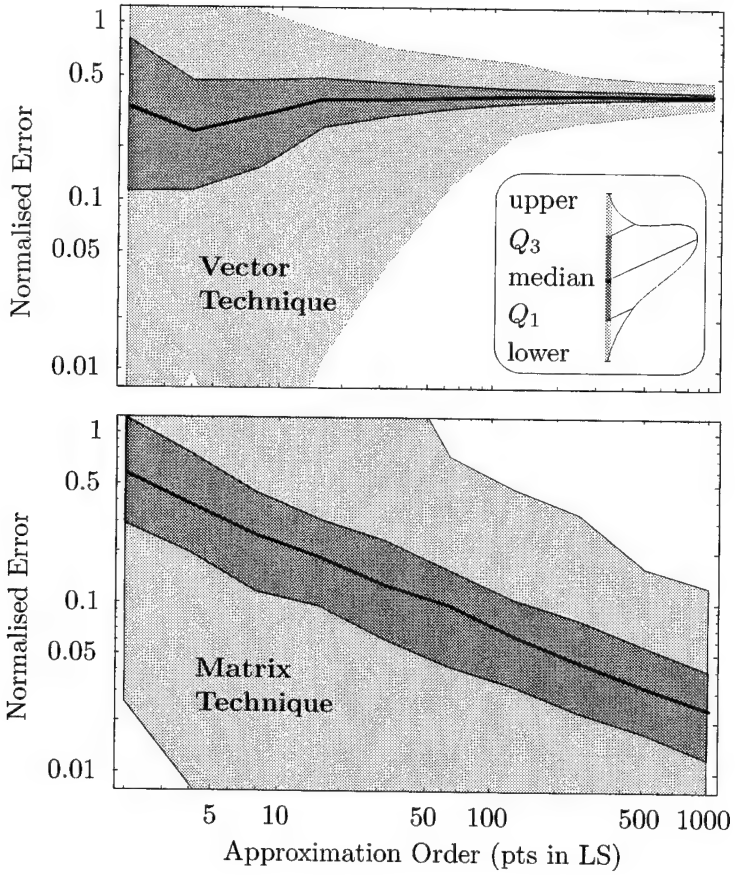


Figure 3.7: Comparison of the normalised errors for the vector and matrix techniques. (Results from an earlier report [25].)

If we were to generate similar distributions for LS approximations with $2, 4, 8, \ldots, 1024$ points, then we would obtain Figure 3.7. Imagine placing Figure 3.6 perpendicular to Figure 3.7 at the 128 point line on the Approximation Order axis, so that the distribution projected out of the page. If we were to repeat this procedure for all the LS distributions using $2, 4, 8, \ldots, 1024$ points and produce a contour plot of the height, then we would

obtain Figure 3.7. The four contour lines represent the lowest outlier, first quartile (Q_1, contains the lowest 25% of the errors), median, third quartile (Q_3, contains the lowest 75% of the errors), and upper outlier of each LS distribution. In other words, Figure 3.7 shows a continuous version of a box-plot. Again each distribution has 1000 sample batches. (For more information on the box-plot, also termed the box-and-whisker plot, see any introductory statistics book; for example Goldman and Weinber [8].)

In Figure 3.7 the vertical axis indicates the normalised error of the LS approximation, while the horizontal axis indicates the order of (or the number of points in) the LS approximation. As we would expect, the matrix technique accuracy improves as we increase the number of data points used to develop the LS approximation. On the contrary, and unexpectedly, the vector technique produces marginally increasing errors as we increase the LS approximation order (except for the change from the 2 to 4 point distributions). Both techniques converge to a solution, but the vector technique converges to an erroneous solution. (We will later show that this incorrect convergence is due to the inconsistency of the LS technique when the measured input data contain noise.)

A slight modification of the vector technique will improve the results. As we noted in an earlier report [25, p. 29] using the median of several low order LS approximations we can improve our estimates of the coefficients a_1 and a_2 in Equation (3.8). The reason we use the median instead of the mean is that we expect some of the LS estimates will be outliers. In fact, from the singular value decomposition theory presented in the appendix of an earlier report [25, § E.1] we saw anecdotal evidence to support the existence of these outliers.

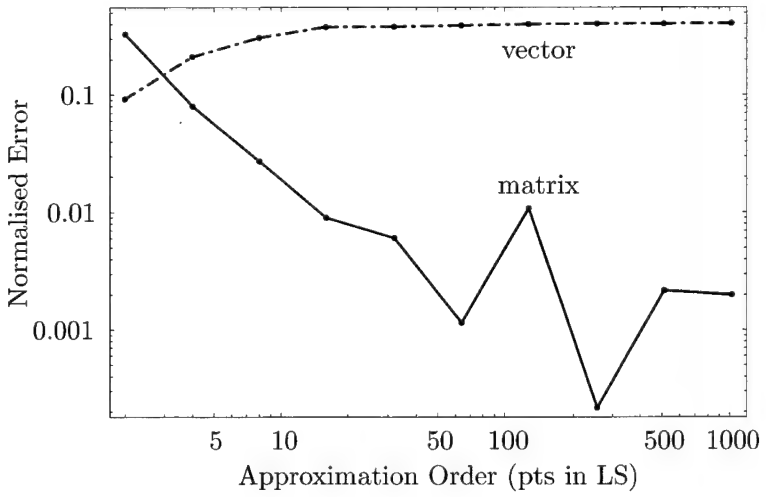


Figure 3.8: The normalised error of the median solutions of both the vector (dashed line) and matrix (solid line) techniques.

Figure 3.8 shows the result of applying this modification (the median of low order LS estimates) to the vector technique. Do not confuse these last two figures. In Figures 3.7 we have plotted the median (along with outliers and quartiles) of the distribution of normalised errors (that is, the *median of the LS errors*). In contrast, in Figure 3.8 we

have plotted the normalised error of the LS coefficient medians (that is, the *error of the LS median*). For example, for a 2 point LS solution, the median of the LS errors is approximately 0.4 (see Figure 3.7) while the error of the LS median is approximately 0.1 (see Figure 3.8).

In Figure 3.8 we see that as the number of points used to develop the LS approximation increases the normalised error also increases. In § 4 we show the reason that the error tails off in Figure 3.8 is that the vector technique returns inconsistent solutions when noise is present in the measurements. Notice that the errors of this modified vector technique are still greater than the errors of the higher order solutions of the matrix technique. Furthermore, this same modification can be applied to the matrix technique, which will subsequently improve the results of the matrix technique (see Figure 3.8).

We reviewed anomalies associated with the development of stress transfer functions (STFs) using simulated strain gauge measurements (superimposed with random noise). Namely, the vector technique was found to deteriorate as the number of points, used to develop the STF, increased. The matrix technique did not suffer this deterioration; in contrast, it improved proportionately to the number of points used to develop the STF. We also found that the both the vector and matrix techniques improved, when we took the median of several solutions of the same order of accuracy.

We now turn out attention to the LS problem, and in the next section (deriving results from scratch) we show how the characteristics of the LS solution can be exploited to our benefit.

4 Least Squares (LS) Fit from a Statistical Perspective

In this section we show why the vector technique produces larger errors than the matrix technique for ill-conditioned problems. We first investigate the simple two-dimensional system, and then generalise these results to the n-dimensional case.

Models of noise in measurements are termed "errors-in-variables" (EIV) models. We cite related work in the EIV models whenever similar results have been published in other fields (especially the field of econometrics).

We begin this section by investigating a simple two-input/single-output linear system, where both the input and output are contaminated by noise.

4.1 LS Approximation with Noise in both Input and Output Measurements

Let's approximate the two-dimensional linear system

$$z(x, y) = ax + by, \tag{4.1}$$

which we term the *true* solution, by a least squares (LS) fit using the m data points $p_i = (\widehat{x}_i, \widehat{y}_i, \widehat{z}_i)$ for $i = 1, 2, \ldots, m$. If the data points p_i were measured (and hence contain noise), then we will get some approximation to Equation (4.1) given by

$$\widehat{z}(x, y) = \widehat{a}\widehat{x} + \widehat{b}\widehat{y}. \tag{4.2}$$

We generalise the above system to an n-dimensional linear system in § 4.6, but for clarity we restrict ourselves to the two-dimensional case shown above for the present.

The system given by Equation (4.1) has a zero intercept, that is, $z(0, 0) = 0$. However, using the substitutions $x = x' - x_0$, $y = y' - y_0$, and $z = z' - z_0$ Equation (4.1) can be easily expanded to the more general system

$$z' - z_0 = a(x' - x_0) + b(y' - y_0)$$

or equivalently

$$z' = c + ax' + by',$$

where $c = z_0 - ax_0 - by_0$, and x', y', and z' all have non-zero means.

Let's decompose the measured inputs $\widehat{x}_i$ and $\widehat{y}_i$, and the measured output $\widehat{z}_i$ into the true inputs, true output, and noise terms. Thus we have that

$$\widehat{x}_i = x_i + u_i, \qquad \widehat{y}_i = y_i + v_i, \qquad \text{and} \qquad \widehat{z}_i = z_i + w_i, \tag{4.3}$$

where x_i and y_i are the true inputs, z_i is the true output, and u_i, v_i, and w_i are randomly distributed noises.

To determine the LS fit we need to define an error function and then minimise it with respect to the sought coefficients $\widehat{a}$ and $\widehat{b}$. Let the error be

$$E^2 = \frac{1}{m}\sum_{i=1}^{m}\left[\widehat{z}(\widehat{x}_i, \widehat{y}_i) - \widehat{z}_i\right]^2.$$

Substituting Equations (4.2) and (4.3) into the above error function yields

$$E^2 = \frac{1}{m}\sum_{i=1}^{m}\Big[\widehat{a}(x_i + u_i) + \widehat{b}(y_i + v_i) - (ax_i + by_i + w_i)\Big]^2. \tag{4.4}$$

Note that E is always positive, unless $\widehat{z}(\widehat{x}_i, \widehat{y}_i) = \widehat{z}_i$ for $i = 1, 2, \ldots, m$ in which case $E = 0$.

Minimising the above error function with respect to both $\widehat{a}$ and $\widehat{b}$ (that is, $dE/d\widehat{a} = 0$ and $dE/d\widehat{b} = 0$) gives that

$$\begin{bmatrix} s_{xx} + 2s_{xu} + s_{uu} & s_{xy} + s_{xv} + s_{yu} + s_{uv} \\ s_{xy} + s_{xv} + s_{yu} + s_{uv} & s_{yy} + 2s_{uv} + s_{vv} \end{bmatrix} \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \begin{bmatrix} a(s_{xx} + s_{xu}) + b(s_{xy} + s_{yu}) + (s_{xw} + s_{uw}) \\ a(s_{xy} + s_{xv}) + b(s_{yy} + s_{yv}) + (s_{yw} + s_{vw}) \end{bmatrix}, \tag{4.5}$$

where s_{ij} is the covariance of the parameters i and j. If $i = j$ then this reduces to the variance or the standard deviation squared, that is $s_{ii} = \sigma_i^2$. For example,

$$s_{xy} = \frac{1}{m}\sum_{i=1}^{m} x_i y_i = \mathrm{cov}(x, y) \qquad \text{and} \qquad s_{uu} = \frac{1}{m}\sum_{i=1}^{m} u_i^2 = \mathrm{var}(u). \tag{4.6}$$

We have developed a linear relationship between the exact coefficients a and b and the approximate coefficients $\widehat{a}$ and $\widehat{b}$ in Equation (4.5). The only assumption we have made so far is that our predictive system is linear. As stated in the introduction this merely requires that the predictive system be a linear function of its variable, and not that the variables be linear in terms of the some "global" variables. For example, the two-dimensional predictive system $z(x, y) = ax + by$ (where a and b are constants) is linear in x and y, even if the functions $x(\xi, \eta)$ and $y(\xi, \eta)$ are highly non-linear in terms of ξ and η.

4.2 Taylor Series Expansion of Noise Terms

Before solving the two-by-two system given by Equation (4.5) we make some approximations. We defined the noise terms u_i, v_i, and w_i as independent, both from each other and from the input and output. However, we also want to investigate the behaviour of the solution when the correlation between the noise terms is small, and thus we proceed with a series expansion analysis.

Let ϵ be the maximum absolute value of any of the s_{ij}, where at least one of either i or j is a noise term u, v, or w, that is

$$\epsilon = \max\Big(|s_{xu}|, |s_{xv}|, |s_{xw}|, |s_{yu}|, |s_{yv}|, |s_{yw}|, |s_{uv}|, |s_{uw}|, |s_{vw}|\Big). \tag{4.7}$$

We can then define the coefficients c_{ij} in terms of ϵ as $c_{ij} = s_{ij}/\epsilon$, and hence by definition $|c_{ij}| \leq 1$. Then using Equation (4.5) and a Taylor's series expansion about $\epsilon = 0$, the coefficient $\widehat{a}$ has the expansion

$$\widehat{a} = \frac{\alpha_0}{\alpha_1} + \epsilon\left(\frac{\alpha_2}{\alpha_1} - 2\frac{\alpha_0\alpha_3}{\alpha_1^2}\right) + \mathcal{O}(\epsilon^2) \tag{4.8}$$

$$= \frac{a\left(s_{xx}s_{yy} - s_{xy}^2\right) + s_{vv}\left(as_{xx} + bs_{xy}\right)}{\left(s_{uu} + s_{xx}\right)\left(s_{vv} + s_{yy}\right) - s_{xy}^2} + \mathcal{O}(\epsilon), \tag{4.9}$$

where

$$\begin{aligned}
\alpha_0 &= a\left(s_{xx}s_{yy} - s_{xy}^2\right) + s_{vv}\left(as_{xx} + bs_{xy}\right), \\
\alpha_1 &= \left(s_{uu} + s_{xx}\right)\left(s_{vv} + s_{yy}\right) - s_{xy}^2, \\
\alpha_2 &= 2ac_{yv}s_{xx} - \left[c_{vw} + a\left(c_{uv} + 2c_{xv} + c_{yu}\right) - bc_{yv} + c_{yw}\right]s_{xy} \\
&\quad - b\left(c_{uv} + c_{xv} + c_{yu}\right)s_{yy} + \left(c_{uw} + ac_{xu} + c_{xw} + bc_{yu}\right)\left(s_{vv} + s_{yy}\right),
\end{aligned}$$

and

$$\alpha_3 = c_{yv}\left(s_{uu} + s_{xx}\right) - \left(c_{uv} + c_{xv} + c_{yu}\right)s_{xy} + c_{xu}\left(s_{vv} + s_{yy}\right).$$

Similarly, the coefficient $\widehat{b}$ has the expansion

$$\begin{aligned}
\widehat{b} &= \frac{\beta_0}{\beta_1} + \epsilon\left(\frac{\beta_2}{\beta_1} - 2\frac{\beta_0\beta_3}{\beta_1^2}\right) + \mathcal{O}(\epsilon^2) \\
&= \frac{b\left(s_{xx}s_{yy} - s_{xy}^2\right) + s_{uu}\left(as_{xy} + bs_{yy}\right)}{\left(s_{uu} + s_{xx}\right)\left(s_{vv} + s_{yy}\right) - s_{xy}^2} + \mathcal{O}(\epsilon),
\end{aligned} \tag{4.10}$$

where

$$\begin{aligned}
\beta_0 &= b\left(s_{xx}s_{yy} - s_{xy}^2\right) + s_{uu}\left(as_{xy} + bs_{yy}\right), \\
\beta_1 &= \left(s_{uu} + s_{xx}\right)\left(s_{vv} + s_{yy}\right) - s_{xy}^2, \\
\beta_2 &= -a\left(c_{uv} + c_{xv} + c_{yu}\right)s_{xx} + \left(c_{vw} + ac_{xv} + bc_{yv} + c_{yw}\right)\left(s_{uu} + s_{xx}\right) \\
&\quad - \left[c_{uw} - ac_{xu} + c_{xw} + b\left(c_{uv} + c_{xv} + 2c_{yu}\right)\right]s_{xy} + 2bc_{xu}s_{yy},
\end{aligned}$$

and

$$\beta_3 = c_{yv}\left(s_{uu} + s_{xx}\right) - \left(c_{uv} + c_{xv} + c_{yu}\right)s_{xy} + c_{xu}\left(s_{vv} + s_{yy}\right).$$

As an important aside, we want to determine how the variance of ϵ decays with the number of points used to develop the least squares (LS) approximation. From Equation (4.7) we know that ϵ is really a covariance between two signals, say f and g, so that $\epsilon = s_{fg} = \mathrm{cov}(f, g)$. From the definition of variance (4.6) we have that

$$\mathrm{var}(\epsilon) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2,$$

but covariance, see Equation (4.6), is defined as

$$\epsilon = \mathrm{cov}(f, g) = \frac{1}{m} \sum_{j=1}^{m} f_j g_j,$$

where n is the number of samples we've taken and m is the number of points in each sample (for simplicity we assume m is the same for all samples). Using these two equations we have that

$$\mathrm{var}(\epsilon) = \frac{1}{m^2 n} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} f_{ij} g_{ij} \right)^2 .$$

The square of a finite sum has the expansion

$$\left(\sum_{j=1}^{m} a_j \right)^2 = \sum_{j=1}^{m} a_j^2 + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} a_j a_k .$$

Hence the variance of ϵ may be expressed as

$$\mathrm{var}(\epsilon) = \frac{1}{m^2} \left[\sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} (f_{ij} g_{ij})^2 + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \frac{1}{n} \sum_{i=1}^{n} f_{ij} g_{ij} f_{ik} g_{ik} \right] .$$

Letting $n \to \infty$ we then have the tidy expression

$$\mathrm{var}(\epsilon) = \frac{1}{m^2} \left[\sum_{j=1}^{m} \mathrm{var}(fg) \right] + 2 \sum_{j=1}^{m-1} \sum_{k=j+1}^{m} \mathcal{E}(fgf'g'),$$

where $\mathcal{E}(x)$ is the expectation (or mean) of x, and both f and f' (and also g and g') are considered different samples of the same function, that is, statistical copies of each other and hence independent.

Now, if f and g are independent then $\mathrm{var}(fg) = \mathrm{var}(f)\,\mathrm{var}(g)$ and $\mathcal{E}(fgf'g') = \mathcal{E}(f)\mathcal{E}(g)\mathcal{E}(f')\mathcal{E}(g')$ (see, for example, Ang and Tang [3]). Since at least one of the variables f or g is a noise term with zero mean, we have that $\mathcal{E}(fgf'g') = 0$, and hence

$$\mathrm{var}(\epsilon) = \frac{\mathrm{var}(f)\,\mathrm{var}(g)}{m},$$

or in terms of standard deviations

$$\sigma_\epsilon = \frac{\sigma_f \sigma_g}{\sqrt{m}}. \tag{4.11}$$

Hence the standard deviation of the term ϵ will reduce to zero as $1/\sqrt{m}$, where m is the number of points used to develop the LS fit.

Lloyd [19] cites Cramér [6] as a source for a similar analysis to that shown above. Lloyd terms the above analysis "sampling moments of a statistic", while Cramér terms it "characteristics of sampling distributions", but unfortunately neither author gives the result shown above. Equation (4.11) is a special case of a result stated (without proof) by

Kendall [14, Eq. (10.21)], who terms the above analysis the "standard errors of bivariate moments".

Lloyd defines the variance with $1/(m-1)$ instead of $1/m$ (as we defined in Equation (4.6)), and gives four good reasons for doing so [19, p. 24]. For m moderately large the difference between these two forms will be negligible, so that this difference will not change the results shown above. The variance, as we've defined it in Equation (4.6), is termed the "maximum-likelihood estimate of variance" [21, p. 417] or the "sample variance" [15, p. 4].

Kendall [15, p. 5–6] provides a useful bias correction procedure (developed by Quenouille) termed the "jack-knife" technique, which is reminiscent of Aitken extrapolation in numerical analysis (see for example Atkinson [4]). Let t_m be a biased estimator of θ, and suppose that t_m has the expectation

$$\mathcal{E}(t_m) = \theta + \sum_{r=1}^{\infty} \frac{a_r}{m^r},$$

where the coefficients a_r are allowed to be functions of θ but not of m. In other words, t_m estimates θ to order $1/m$ (that is, $\mathcal{E}(t_m) = \theta + \mathcal{O}(1/m)$).

From the m observations we can calculate the estimate t_{m-1}; in fact, we can calculate m of these t_{m-1} estimates (since there are m possible subsets of $(m-1)$ observations). Let $\bar{t}_{m-1}$ denote the mean of these m estimates of t_{m-1}; that is, $\bar{t}_{m-1} = \frac{1}{m}\sum_{i=1}^{m}(t_{m-1})_i$. A new estimate of θ is

$$t'_m = m\,t_m - (m-1)\,\bar{t}_{m-1}, \tag{4.12}$$

where $\mathcal{E}(t'_m) = \theta + \mathcal{O}(1/m^2)$ and hence t'_m is a second order (in m) approximation of θ. This jack-knife principle can be continued recursively (see Kendall), we give the next estimate

$$t''_m = \frac{m^2\,t'_m - (m-1)^2\,\bar{t}'_{m-1}}{m^2 - (m-1)^2}, \tag{4.13}$$

where $\mathcal{E}(t''_m) = \theta + \mathcal{O}(1/m^3)$ shows that t''_m is a third order (in m) approximation of θ.

Since Equations (4.9) and (4.10) are biased to order $1/m$, remember that $\epsilon = \mathcal{O}(1/m)$, we can use the jack-knife bias correction technique. We will later show, however, that not only are the estimates for the coefficients given by Equations (4.9) and (4.10) biased, they are also inconsistent. (In fact, we already suspect this inconsistent behaviour from the results of the vector technique shown in Figure 3.7 on page 18.)

Under the assumption of uncorrelated noise we derived approximations relating the exact and approximate coefficients of our linear predictive system (Equations (4.9) and (4.10)). From these expressions we see that as the two inputs x and y become more correlated the denominators in Equations (4.9) and (4.10) tend to zero; this problem is exacerbated by noise in the signals. Thus our coefficient estimates $\widehat{a}$ and $\widehat{b}$ become worse as the system becomes more ill-conditioned (even when a small amount of noise is present). We have also seen that the variance of the bias ϵ is inversely proportional to the number of points used to develop the stress transfer function (Equation (4.11)). Finally, we learned of a general bias correction procedure termed the jack-knife technique.

4.3 Input and Output Noise Independent

In this section we assume that all noise terms are uncorrelated with each other and uncorrelated with the true signals, that is, $\epsilon = 0$.

Introduce the *noise terms* as

$$\eta_x = \frac{s_{uu}}{s_{xx}} \quad \text{and} \quad \eta_y = \frac{s_{vv}}{s_{yy}}, \tag{4.14}$$

and the *correlation coefficient* between the two inputs x and y as

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}. \tag{4.15}$$

The variables η_x and η_y may be thought of as the noise-to-signal ratio in the input signals x and y respectively, and note that $\eta_x, \eta_y \geq 0$. Furthermore, the Cauchy inequality [1] states that $\left(\sum_{i=1}^n x_i y_i\right)^2 \leq \sum_{i=1}^n x_i^2 \sum_{j=1}^n y_j^2$, with the equality holding only if $x_i = cy_i$ for some constant c, and thus $-1 \leq r \leq 1$.

If $\epsilon = 0$ then the coefficient $\widehat{a}$ in Equation (4.9) becomes

$$\begin{aligned}\widehat{a} &= a\left[\frac{(1+\eta_y) - r^2 + \eta_y r \frac{b}{a}\sqrt{\frac{s_{yy}}{s_{xx}}}}{(1+\eta_x)(1+\eta_y) - r^2}\right] \\ &= a\left[1 - \eta_x \frac{1 + \eta_y - r(b\eta_y\sqrt{s_{yy}})/(a\eta_x\sqrt{s_{xx}})}{(1+\eta_x)(1+\eta_y) - r^2}\right].\end{aligned} \tag{4.16}$$

Similarly $\widehat{b}$ can be written as

$$\widehat{b} = b\left[1 - \eta_y \frac{1 + \eta_x - r(a\eta_x\sqrt{s_{xx}})/(b\eta_y\sqrt{s_{yy}})}{(1+\eta_x)(1+\eta_y) - r^2}\right]. \tag{4.17}$$

Using a more generalised approach (the assumption of independent noise was not used), Schneeweiß [30] develops the matrix form of the above expression.

The two terms $a\sqrt{s_{xx}}$ and $b\sqrt{s_{yy}}$ measure the importance of a particular parameter in the linear relationship, since $a\sqrt{s_{xx}}$ includes the spread of the true signal x and the coefficient a. In fact these two terms are really just a scaling of the original equation

$$\begin{aligned}z &= ax + by \\ &= \left(a\sqrt{s_{xx}}\right)\frac{x}{\sqrt{s_{xx}}} + \left(b\sqrt{s_{yy}}\right)\frac{y}{\sqrt{s_{yy}}}\end{aligned}$$

where $x/\sqrt{s_{xx}}$ and $y/\sqrt{s_{xx}}$ are non-dimensionalised versions of x and y respectively (that is, they've been scaled by their respective standard deviations).

Making some simplifying assumptions we get the underlying form of the above expressions for $\widehat{a}$. Assume that the two noise-to-signal terms are equal (that is $\eta_x = \eta_y$) and that the variance of the two true input signals are also equal (that is, $s_{xx} = s_{yy}$). Then we can simplify the above expression for $\widehat{a}$ to

$$\widehat{a} = a\left[1 - \eta\frac{(1 + \eta - \frac{b}{a}r)}{(1+\eta+r)(1+\eta-r)}\right]. \tag{4.18}$$

where $\eta = \eta_x = \eta_y$ is a noise-to-signal term. (The coefficient $\widehat{b}$ can be similarly simplified.) We can re-arrange the above equation to determine the *relative error* of the approximate coefficient $\widehat{a}$ as

$$\frac{\widehat{a}}{a} - 1 = -\eta \frac{(1 + \eta - \frac{b}{a} r)}{(1 + \eta + r)(1 + \eta - r)}. \tag{4.19}$$

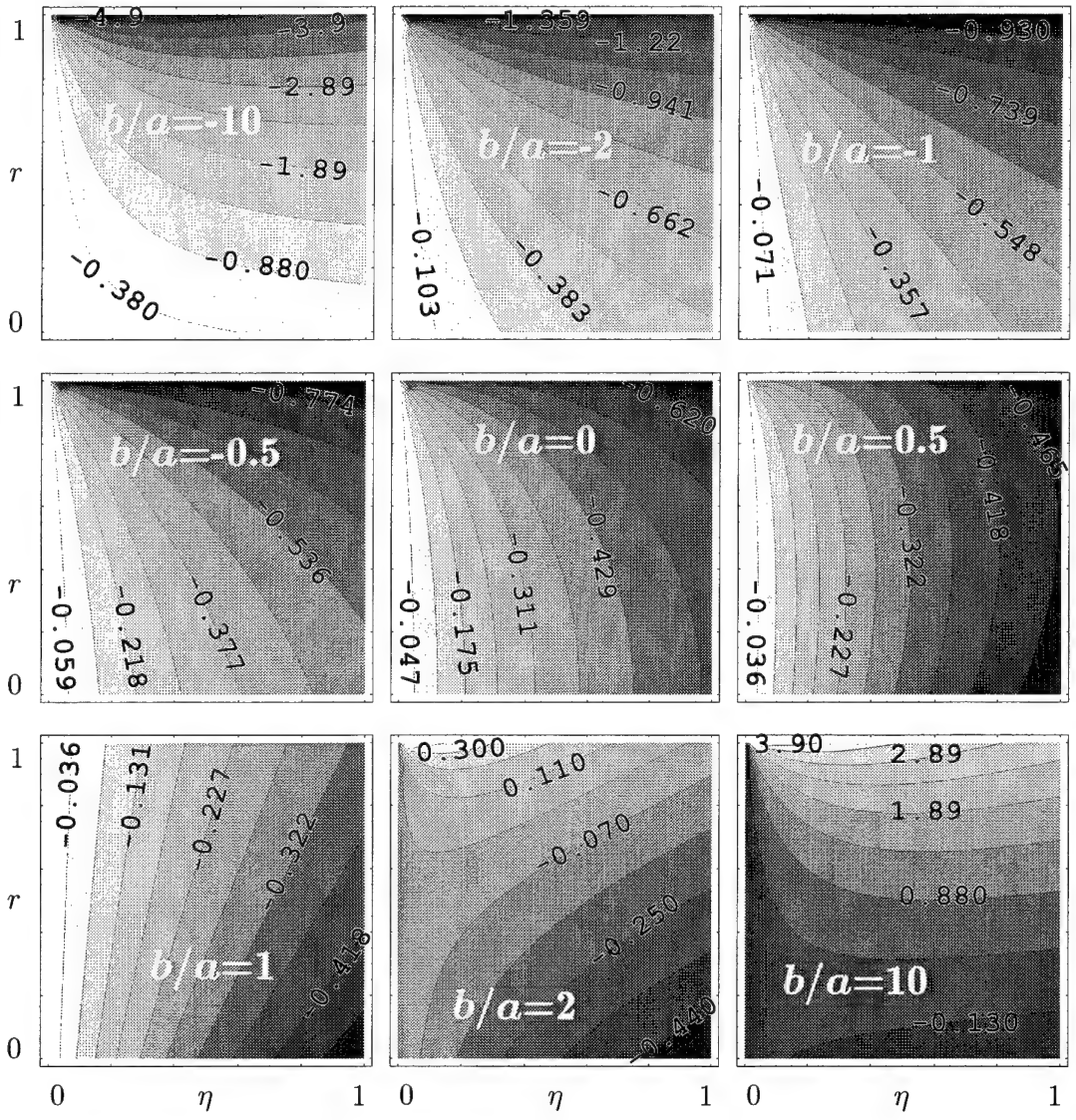


Figure 4.1: The relative error of the approximate coefficient $\widehat{a}$ plotted against correlation r and noise η for several values of the ratio of exact coefficients b/a.

Figure 4.1 shows contour plots of the relative error of the approximate coefficient $\widehat{a}$ for several values of b/a (the ratio of exact coefficients). Notice that the relative error is

not symmetric about $b/a = 0$, that is negative values of b/a produce different results to positive values. Also note from Equation (4.19) that the approximated coefficient matches the exact coefficient ($\widehat{a} = a$) only if $\eta = 0$ or $\eta = (b/a)r - 1$. The second condition, $\eta = (b/a)r - 1$, suggests that under some conditions the effects of noise exactly cancel the effects of correlated inputs to yield the exact solution!

If the coefficients of the two inputs have the same magnitude ($|a| = |b|$), then from Equation (4.18) the numerator and one of the terms in the denominator of the noise amplification term cancel, and hence we have that

$$\widehat{a} = a\left[1 - \eta\frac{1}{1+\eta+|r|}\right] \qquad (\text{if } |a| = |b|). \tag{4.20}$$

The modulus of the correlation (that is, $|r|$) was needed in the above equation for the following reason. First note that the term $r\ b/a$ is always positive, since if the coefficients have different signs ($b/a < 0$), then the correlation must also be negative ($r < 0$). Now if r is positive then the modulus operator has no effect, so we need only consider the case $r < 0$. If the correlation is negative ($r < 0$) then the numerator $1 + \eta - rb/a$ cancels out the denominator term $1 + \eta + r$. Hence the denominator can be expressed as $1 + \eta + |r|$.

From the above equation we see that input correlation is no longer a problem; in fact, it's beneficial. This last statement assumes ϵ is zero; that is all noise terms are uncorrelated both to each other and to the true signals. If we have ϵ small but non-zero (that is, $0 < \epsilon \ll 1$), then we need to temper the above "beneficial" statement with the restriction that $|r|$ is not too large. The reason for this restriction on r is that the coefficient of ϵ is proportional to $1/\alpha_1^2 = 1/(1+\eta-r)^2$ (where we have used the simplification $\eta = \eta_x = \eta_y$ in Equation (4.8)). Hence for $0 < \epsilon \ll 1$ our assumption that the ϵ term in Equation (4.8) remains small is violated as $r \to 1$.

We now see, from Equation (4.20), that the least squares (LS) solution will always under-estimate the coefficient a whenever there is some noise in the input signal. (This result is well known for the case of zero correlation $r = 0$ and is termed the "errors-in-variables" model.) In other words, the slope of the estimate $\widehat{a}$ is always closer to zero than the true slope a. This under-estimation is termed "attenuation bias" by Johnston and DiNardo [13].

At first it appears from Equation (4.16) that scaling the inputs x and y might improve $\widehat{a}$ (the estimation of the coefficient). However, using the notation $x' = x/c_x$ and $y' = y/c_y$ (where c_x and c_y are constants), we can show that scaling has no effect on the solution of the coefficients. Let a' and b' be the associated scaled coefficients a and b respectively, then from the relation $ax = a'x'$ we must have $a' = ac_x$, and similarly $b' = bc_y$. We now see that $a\sqrt{s_{xx}} = a'\sqrt{s_{x'x'}}$ and hence, from Equation (4.16), scaling the inputs x and y has no effect on estimating $\widehat{a}$.

For η small, or more precisely $\eta^2 \ll 1 - r^2$, we have the approximation

$$\widehat{a} \approx a\left[1 - \eta\frac{\left(1 - \frac{b}{a}r\right)}{(1+r)(1-r)}\right] \qquad (\text{if } \eta^2 \ll 1 - r^2).$$

We can think of the term multiplying η as the *noise amplification function*. For example, if we have 10% noise in the input measurements and the noise amplification function has a

value of 3, then the approximate coefficient $\widehat{a}$ would have a relative error of $3 \times 10\% = 30\%$. The restriction $\eta^2 \ll 1 - r^2$ implies that as the two inputs x and y become more correlated the noise-to-signal ratio must decrease by at least a proportional amount if the small noise assumption is to be used.

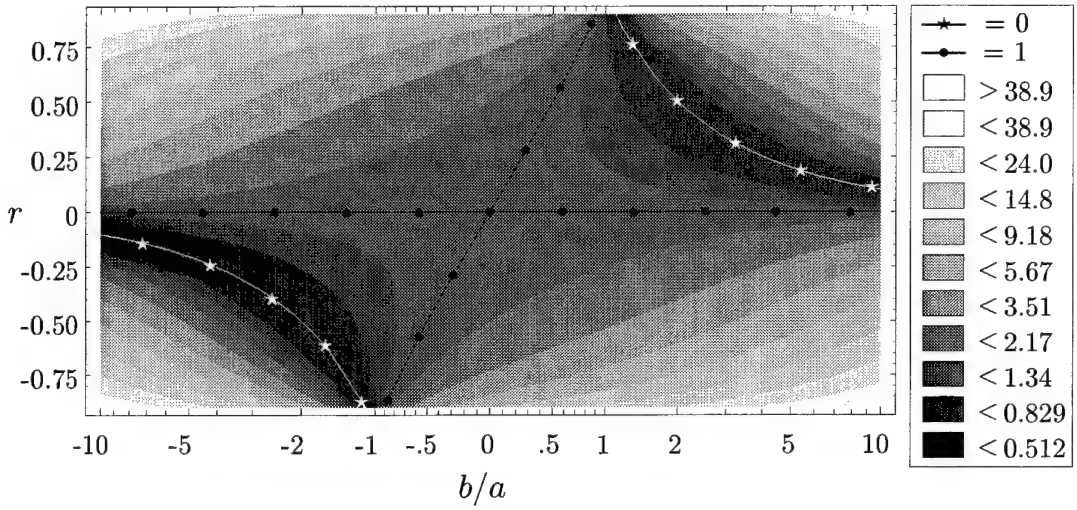


Figure 4.2: Absolute value of the noise amplifying function, $|(1 - rb/a)/(1 - r^2)|$, plotted against correlation r and ratio of exact coefficients b/a, assuming a small amount of noise ($\eta^2 \ll 1 - r^2$). The white line highlights the location of the zero noise amplification line. The black line denotes where input and output noise are equal.

Figure 4.2 shows a contour plot of the noise amplification function plotted against the ratio of exact coefficients b/a and the correlation r. In fact, this contour plot shows the logarithm of the absolute value of the noise amplification, that is $\log\left|(1 - rb/a)/(1 - r^2)\right|$, which explains the exponential scale on the contour legend. Finally, note that the ticks marks on the b/a-axis are neither linear nor logarithmic. The ticks were in fact spaced out using an inverse hyperbolic sine function, $\sinh^{-1}(x)$, so that for small values of x the ticks are spaced almost linearly, while for $|x|$ large the spacing becomes logarithmic.

The minimum noise amplification is shown as a white line $r = 1/(b/a)$. For $b/a > 0$ and $b/a < 0$ all contours above and below this line respectively are negative. The black lines ($r = 0$ and $r = b/a$) denote the locations where input and output noise are equal, that is unit amplification. Again notice that the noise amplification function is not symmetric about $b/a = 0$, that is negative values of b/a produce different results to positive values. However, the noise amplification is anti-symmetric about $r = -b/a$.

In this section we ignored the bias ϵ, since we can make it as small as we like (owing to the fact that $\epsilon = \mathcal{O}(1/m)$) simply by using more points. Thus ignoring bias we re-wrote the relation between the approximate and exact coefficients ($\widehat{a}$ and a respectively) in terms of noise-to-signal ratios (η_x and η_y), correlation (r), and importance weightings ($a\sqrt{s_{xx}}$ and $b\sqrt{s_{yy}}$). These relations are given by Equations (4.16) and (4.17) for the approximate

coefficients $\widehat{a}$ and $\widehat{b}$ respectively.

To determine the form of these expressions we further assumed that the two input signals (x and y) and associated noises (u and v) had the same variance. We subsequently found that the expression relating the approximate and exact coefficients, Equation (4.18), was a function of the noise to signal ratio (η), correlation (r), and ratio of exact coefficients (b/a). Our approximate coefficient was identical to the exact coefficient (that is, we obtain zero error) under two conditions; when the noise is zero ($\eta = 0$) or when the noise exactly cancels the correlation ($\eta = (b/a)r - 1$, which was a surprising result). Contour plots of the approximate coefficient's relative error function, Figure 4.1, reinforced these results.

We found that if the exact coefficients had the same magnitude ($|a| = |b|$), then correlation was no longer a problem, even for ill-conditioned systems. On the contrary, correlation was now beneficial. The well known result of LS under-prediction (from errors-in-variables theory) was observed in this simplified form (Equation (4.20)).

Assuming, instead, that the noise was much closer to zero than the correlation was to unity ($\eta^2 \ll 1 - r^2$), we gained a greater understanding of the effects of exact coefficient ratio (b/a) on errors. A contour plot of the noise amplifying function, Figure 4.2, emphasised the importance of having input signals of similar order of influence on the output signal (that is, $b/a \approx 1$). From this same plot we discovered when the error in coefficient estimation was zero, and also under what conditions the noise would be amplified detrimentally.

4.4 Correcting for the LS Inconsistency

In this subsection we approximate the true least squares (LS) coefficients a and b using noise estimates, input correlation estimates, and the approximate LS coefficients $\widehat{a}$ and $\widehat{b}$. We will first develop results for the simple case of equal noise (in all inputs), and then generalise this result. Before beginning however, it is worthwhile clarifying the terms "biased" and "consistent".

Kendall [15, p. 3] defines consistency in the following manner. An estimator $\widehat{\vartheta}_n$, computed from a sample of n values, is said to be *consistent* if, for any positive μ and ν, there exists some N such that the probability that $|\widehat{\vartheta}_n - \vartheta| < \mu$ is greater than $1 - \nu$ for all $n > N$. In the notation of the theory of probability,

$$P\left\{|\widehat{\vartheta}_n - \vartheta| < \mu\right\} > 1 - \nu, \qquad \text{for all } n > N.$$

In other words, given any small quantity μ we can find a large enough sample size such that, for all samples over that size, the probability that $\widehat{\vartheta}$ differs from the true value by more than μ is as near zero as we please. (At least one author [16] appears to use the terms "inconsistent" and "asymptotic bias" interchangeably.)

If we require that for all n the mean value of $\widehat{\vartheta}$ shall be ϑ, then

$$\mathcal{E}(\widehat{\vartheta}_n) = \vartheta \qquad \text{for all } n$$

defines an *unbiased estimator*, where $\mathcal{E}(\cdot)$ denotes the expectation (or mean). Kendall states that the median or mode, instead of the mean, can also be chosen in defining

an "unbiased" estimator. Note that an estimator may be both consistent and unbiased, consistent and biased, or inconsistent and unbiased; that is, one does not imply the other.

We reproduce Equation (4.18) below

$$\widehat{a} = a\left[1 - \eta\frac{\left(1+\eta-\frac{b}{a}r\right)}{(1+\eta+r)(1+\eta-r)}\right] + \mathcal{O}(\epsilon),$$

and similarly for $\widehat{b}$ we have the expression

$$\widehat{b} = b\left[1 - \eta\frac{\left(1+\eta-\frac{a}{b}r\right)}{(1+\eta+r)(1+\eta-r)}\right] + \mathcal{O}(\epsilon).$$

The above two equations express $\widehat{a}$ and $\widehat{b}$ in terms of a and b. Thus re-arranging we can solve for a and b in terms of $\widehat{a}$ and $\widehat{b}$ giving (after a great deal of algebraic manipulation)

$$a = \widehat{a}\left[1 + \frac{\eta}{(1+r)}\frac{\left(1-\frac{\widehat{b}}{\widehat{a}}r\right)}{(1-r)}\right] + \mathcal{O}(\epsilon).$$

We obtain a similar expression for b. To use the above consistency correction formula we need estimates of the correlation r and the amount of noise in the signal η.

If we perform the same analysis on the more general form of the coefficients $\widehat{a}$ and $\widehat{b}$ given by Equations (4.16) and (4.17) we obtain

$$a = \widehat{a}\left[1 + \frac{s_{uu}}{s_{xx}}\frac{\left(1-\frac{s_{vv}}{s_{uu}}\sqrt{\frac{s_{xx}}{s_{yy}}}\frac{\widehat{b}}{\widehat{a}}r\right)}{(1-r^2)}\right] + \mathcal{O}(\epsilon) \tag{4.21}$$

$$= \widehat{a}\left[1 + \frac{\eta_x}{(1+r)}\frac{(1-\phi r)}{(1-r)}\right] + \mathcal{O}(\epsilon), \tag{4.22}$$

where η_x, η_y, and r are defined by Equations (4.14) and (4.15) respectively, and

$$\phi = \frac{\eta_y}{\eta_x}\frac{\widehat{b}\sqrt{s_{yy}}}{\widehat{a}\sqrt{s_{xx}}}.$$

Remember from an earlier argument that $a\sqrt{s_{xx}}$ and $b\sqrt{s_{yy}}$ may be thought of as importance weightings of the inputs (x and y) in determining the output z. Hence the above ϕ term may be though of as the product of noise-term-ratio (η_y/η_x) and importance-weighting-ratio ($\widehat{b}\sqrt{s_{yy}}/(\widehat{a}\sqrt{s_{xx}})$). Making the necessary changes, the coefficient b has the same form as the above equation for a.

For the above consistency correction formula to be of any use, we must have estimates of the noise variances (s_{uu} and s_{vv}) and the true input signal variances (s_{xx} and s_{yy}). Obviously the accuracy of these variance estimates will reflect the accuracy of the LS consistency correction. We now give estimates of these variances.

Remembering that measurements of true values are denoted by a hatted notation, for example $\hat{x}_i = x_i + u_i$ (see Equation (4.3)), and using the definition of variance given by Equation (4.6) we know that

$$\begin{aligned} s_{\hat{x}\hat{x}} &= \frac{1}{m}\sum_{i=1}^{m} \hat{x}_i^2 \\ &= \frac{1}{m}\sum_{i=1}^{m} (x_i + u_i)^2 \\ &= \frac{1}{m}\sum_{i=1}^{m} x_i^2 + 2\frac{1}{m}\sum_{i=1}^{m} x_i u_i + \frac{1}{m}\sum_{i=1}^{m} u_i^2 \\ &= s_{xx} + 2s_{xu} + s_{uu}. \end{aligned}$$

We know, however, that the covariance between a signal and noise is less than or equal to ϵ (see Equation (4.7)), and hence $s_{xu} = \mathcal{O}(\epsilon)$. An estimate of the variance of the true signal x is then given by

$$s_{xx} = s_{\hat{x}\hat{x}} - s_{uu} + \mathcal{O}(\epsilon), \tag{4.23}$$

and hence we must have an estimate of the input noise variance s_{uu}. One way to obtain this s_{uu} estimate is to take repeated measurement of x with all input variables (including x) fixed.

We can obtain an estimate of r, the correlation coefficient between the two inputs, in a similar fashion. Using Equation (4.7) we obtain that $s_{\hat{x}\hat{y}} = s_{xy} + \mathcal{O}(\epsilon)$; then substituting this result and Equation (4.23) into Equation (4.15) gives an estimate of the correlation coefficient

$$r = \frac{s_{\hat{x}\hat{y}}}{\sqrt{(s_{\hat{x}\hat{x}} - s_{uu})(s_{\hat{y}\hat{y}} - s_{vv})}} + \mathcal{O}(\epsilon), \tag{4.24}$$

where we have again made use of the input noise variances s_{uu} and s_{vv}.

Substituting the estimates for variances s_{xx} and s_{yy} (Equation (4.23)) and for the correlation coefficient (Equation (4.24)) into the consistency correction formula given by Equation (4.21) gives

$$a = \widehat{a}\left[1 + \frac{s_{uu}(s_{\hat{y}\hat{y}} - s_{vv}) - s_{vv}\,\frac{\widehat{b}}{\widehat{a}}\,s_{\hat{x}\hat{y}}}{(s_{\hat{x}\hat{x}} - s_{uu})(s_{\hat{y}\hat{y}} - s_{vv}) - s_{\hat{x}\hat{y}}^2}\right] + \mathcal{O}(\epsilon). \tag{4.25}$$

Notice that the above equation has the same form as Equation (4.21) when re-arranged. The above form (in contrast to the form given in Equation (4.21)) was chosen in order to simplify the series expansion analysis that will be undertaken below.

Equation (4.25) represents the consistency correction for $\widehat{a}$ using $s_{\hat{x}\hat{x}}$ and $s_{\hat{y}\hat{y}}$ (estimates of the true input variations), and $s_{\hat{x}\hat{y}}$ (an estimate of the covariance of the true inputs). These estimates, however, use the exact noise variances s_{uu} and s_{vv}. What if we have to "guesstimate" these variances themselves? Below we perform a sensitivity analysis on the estimates of the noise variances.

4.4.1 Sensitivity of LS Correction to Errors in Noise Estimation

Let us estimate the true noise variance s_{uu} and s_{vv} by $\widehat{s}_{uu}$ and $\widehat{s}_{vv}$ respectively. (Notice that we have hatted the variance symbol s instead of the noise terms u and v, since $\hat{u}$ and $\hat{v}$ would imply exact measurements of the noises.) We can then write down a relative relation between these true and estimated variances as

$$\widehat{s}_{uu} = s_{uu}(1+\delta_u) \qquad \text{and} \qquad \widehat{s}_{vv} = s_{vv}(1+\delta_v),$$

where δ_u and δ_v are the relative errors in our estimates of the noise variance s_{uu} and s_{vv} respectively. We see that δ_u, for example, really does represent relative error upon re-arranging the above equation to give $\delta_u = (\widehat{s}_{uu} - s_{uu})/s_{uu}$.

Substituting the above two equations into the consistency correction formula given by Equation (4.25), and re-arranging, gives that

$$a = \widehat{a}\left[1 + \widehat{\eta}_x \frac{(1-\widehat{\phi}\,\widehat{r}) - (\widehat{\phi}\,\widehat{r})\delta_u + (1+\widehat{\eta}_y)\delta_v}{(1-\widehat{r}^{\,2}) + (1-\widehat{r}^{\,2}+\widehat{\eta}_x)\delta_u + (1-\widehat{r}^{\,2}+\widehat{\eta}_y)\delta_v + (1-\widehat{r}^{\,2}+\widehat{\eta}_x+\widehat{\eta}_y+\widehat{\eta}_x\widehat{\eta}_y)\delta_u\delta_v}\right] + \mathcal{O}(\epsilon), \quad (4.26)$$

where

$$\widehat{\eta}_x = \frac{\widehat{s}_{uu}}{s_{\hat{x}\hat{x}} - \widehat{s}_{uu}}, \qquad \widehat{\eta}_y = \frac{\widehat{s}_{vv}}{s_{\hat{y}\hat{y}} - \widehat{s}_{vv}},$$

$$\widehat{r} = \frac{s_{\hat{x}\hat{y}}}{\sqrt{(s_{\hat{x}\hat{x}} - \widehat{s}_{uu})(s_{\hat{y}\hat{y}} - \widehat{s}_{vv})}}, \qquad \text{and} \qquad \widehat{\phi} = \frac{\widehat{\eta}_y}{\widehat{\eta}_x}\frac{\widehat{a}\sqrt{s_{\hat{y}\hat{y}} - \widehat{s}_{vv}}}{\widehat{b}\sqrt{s_{\hat{x}\hat{x}} - \widehat{s}_{uu}}}.$$

Performing a Taylor's series expansion of Equation (4.26) about the relative errors $\delta_u = 0$ and $\delta_v = 0$ gives

$$a = \widehat{a}\left(1 + \frac{\widehat{\eta}_x}{(1-\widehat{r}^{\,2})}\left\{(1-\widehat{\phi}\,\widehat{r}) - \delta_u\left[1 + \widehat{\eta}_x\frac{(1-\widehat{\phi}\,\widehat{r})}{(1-\widehat{r}^{\,2})}\right] + \delta_v\widehat{r}\left[\widehat{\phi} + \widehat{\eta}_y\frac{(\widehat{\phi}-\widehat{r})}{(1-\widehat{r}^{\,2})}\right]\right\}\right) + \mathcal{O}(\delta_u) + \mathcal{O}(\delta_v) + \mathcal{O}(\delta_u\delta_v) + \mathcal{O}(\epsilon). \quad (4.27)$$

Again, note the significance of the special case where the importance weightings are equal (that is, $\widehat{\phi} = 1$), for which the above equation becomes

$$a = \widehat{a}\left(1 + \frac{\widehat{\eta}_x}{(1-\widehat{r}^{\,2})}\left\{(1-\widehat{r}) - \delta_u\left[1 + \frac{\widehat{\eta}_x}{(1+\widehat{r})}\right] + \delta_v\widehat{r}\left[1 + \frac{\widehat{\eta}_y}{(1+\widehat{r})}\right]\right\}\right) + \mathcal{O}(\delta_u) + \mathcal{O}(\delta_v) + \mathcal{O}(\delta_u\delta_v) + \mathcal{O}(\epsilon). \quad (4.28)$$

We can now clearly see that if $|\delta_u|, |\delta_u| \ll |1-\widehat{r}|$, then our estimate of the exact coefficient is a good one. Notice that, unlike Equation (4.27), the factors multiplying δ_u and δ_v in Equation (4.28) are bounded as $\widehat{r}$ tends to unity.

The large factors multiplying δ_u and δ_v (large as $\widehat{r} \to 1$) originate from the denominator of Equation (4.26). In fact, our least squares (LS) correction becomes singular when this denominator is zero. Writing Equation (4.26) in terms of δ_u gives

$$a = \widehat{a}\left[1 + c_0\frac{(c_1 - \delta_u)}{(c_2 - \delta_u)}\right] + \mathcal{O}(\epsilon), \quad (4.29)$$

where

$$c_0 = -\frac{\widehat{\eta}_x\,\widehat{\phi}\,\widehat{r}}{(1-\widehat{r}^{\,2}+\widehat{\eta}_x)},$$

$$c_1 = \frac{(1-\widehat{\phi}\,\widehat{r})+(1+\widehat{\eta}_y)\delta_v}{\widehat{\phi}\,\widehat{r}}, \qquad \text{and}$$

$$c_2 = -\frac{(1-\widehat{r}^{\,2})+(1-\widehat{r}^{\,2}+\widehat{\eta}_y)\delta_v}{(1-\widehat{r}^{\,2}+\widehat{\eta}_x)+(1-\widehat{r}^{\,2}+\widehat{\eta}_x+\widehat{\eta}_y+\widehat{\eta}_x\widehat{\eta}_y)\delta_v}.$$

Similar expressions arise if we re-write Equation (4.26) in terms of δ_v instead of δ_u.

Figure 4.3 shows the general form of Equation (4.29). Note the singularity at $\delta_u = c_2$, and that the sign of c_0 determines whether $a/\widehat{a}$ tends to negative or positive infinity ($c_0 < 0$ or $c_0 > 0$ respectively) as δ_u approaches the singularity from the positive side ($\delta_u \to c_2^+$).

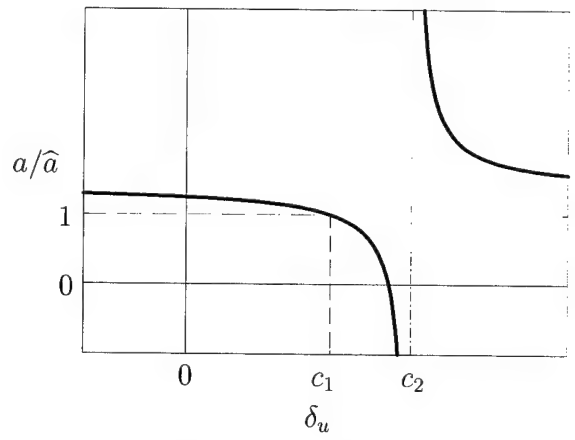


Figure 4.3: Sensitivity of the corrected LS solution to errors in the noise estimate, where $a/\widehat{a} = 1 + c_0(c_1 - \delta_u)/(c_2 - \delta_u) + \mathcal{O}(\epsilon)$. The sign of c_0 determines whether the form shown ($c_0 > 0$) or the inverted form ($c_0 < 0$) is produced.

Let's investigate the conditions under which the LS correction is relatively insensitive to errors in our noise estimate. We have just seen that under some conditions bad noise estimates have the potential to send our LS correction to infinity. Consider the form of the Taylor's series expansion of corrected LS to noise estimate sensitivity, that is Equation (4.27). We want to determine when the factors multiplying δ_u and δ_v remain bounded, since under this bounded condition our corrected LS is insensitive to errors in noise estimates. The last statement is true provided $|\delta_u|, |\delta_u| \ll |1 - \widehat{\phi}\,\widehat{r}\,|$, since we essentially want to form an argument that says we can ignore the terms δ_u and δ_v and their multiplying factors in Equation (4.27).

From Equation (4.27) the factor multiplying δ_u is (ignoring the addition of unity since

it is bounded)

$$\widehat{\eta}_x \frac{1-\widehat{\phi}\,\widehat{r}}{1-\widehat{r}^{\,2}} = \frac{1}{(1+\widehat{r})} \frac{\widehat{\eta}_x\,(1-\widehat{\phi}\,\widehat{r})}{(1-\widehat{r})}.$$

The first fraction on the right hand side is bounded as $\widehat{r} \rightarrow 1$, so we need only consider the second fraction on the right hand side. We want to determine when this fraction is bounded; to be pragmatic let's choose that bound to be unity. We then have that the condition for boundedness is

$$|1-\widehat{\phi}\,\widehat{r}\,| \leq \frac{1}{\widehat{\eta}_x}|1-\widehat{r}|.$$

It can be shown that the above condition is satisfied by either of the two conditions

$$\frac{1-(1-\widehat{r})/\widehat{\eta}_x}{\widehat{r}} \leq \widehat{\phi} \leq \frac{1+(1-\widehat{r})/\widehat{\eta}_x}{\widehat{r}} \quad (\text{for } \widehat{r} \leq 1) \tag{4.30}$$

or

$$\frac{1+(1-\widehat{r})/\widehat{\eta}_x}{\widehat{r}} \leq \widehat{\phi} \leq \frac{1-(1-\widehat{r})/\widehat{\eta}_x}{\widehat{r}} \quad (\text{for } r \geq 1). \tag{4.31}$$

Similarly the factor multiplying δ_v in Equation (4.27) (ignoring $\widehat{r}$ and $\widehat{\phi}$ since these terms are at least bounded, if not near unity) is given by

$$\widehat{\eta}_y \frac{\widehat{\phi}-\widehat{r}}{1-\widehat{r}^{\,2}} = \frac{\widehat{\phi}}{(1+\widehat{r})} \frac{\widehat{\eta}_y\,(1-\widehat{r}/\widehat{\phi})}{(1-\widehat{r}\,)}.$$

Again, we need only consider the second fraction on the right hand side, which is bounded by unity under either of the two conditions

$$\frac{\widehat{r}}{1+(1-\widehat{r})/\widehat{\eta}_y} \leq \widehat{\phi} \leq \frac{\widehat{r}}{1-(1-\widehat{r})/\widehat{\eta}_y} \quad (\text{for } r \leq 1), \tag{4.32}$$

or

$$\frac{\widehat{r}}{1-(1-\widehat{r})/\widehat{\eta}_y} \leq \widehat{\phi} \leq \frac{\widehat{r}}{1+(1-\widehat{r})/\widehat{\eta}_y} \quad (\text{for } r \geq 1). \tag{4.33}$$

The conditions given by Equations (4.30)–(4.33) are mutually exclusive (except for the end-points) if $\widehat{\eta}_x, \widehat{\eta}_y \geq 1$, as we would expect. This last statement simply means that for a large amount of noise our correction to the LS solution is sensitive to errors in our noise estimation. Any value of the importance weighting $\widehat{\phi}$ that satisfies either of the conditions given by Equations (4.30) or (4.31) as well as either of the conditions given by Equations (4.32) or (4.33) guarantees that the corrected LS is insensitive to errors in noise estimation; this follows since we are essentially showing that the ratio $a/\widehat{a}$ is on the flatter part of Figure 4.3. As a final point, note that these four conditions on $\widehat{\phi}$ become harder and harder to satisfy as the approximate correlation $\widehat{r}$ tends to unity; again something we would expect since the measurement data set is becoming more ill-conditioned.

Figure 4.4 shows how the normalised error (in determining the coefficients a and b) varies with errors in noise estimation δ_u and δ_v. Both noise estimation errors were varied simultaneously, that is, $\delta = \delta_u = \delta_v$, which explains why the minimum normalised error is not zero. The lines shown in Figure 4.4 merely connect the associated points, hence the jagged appearance of the lines near the minima.

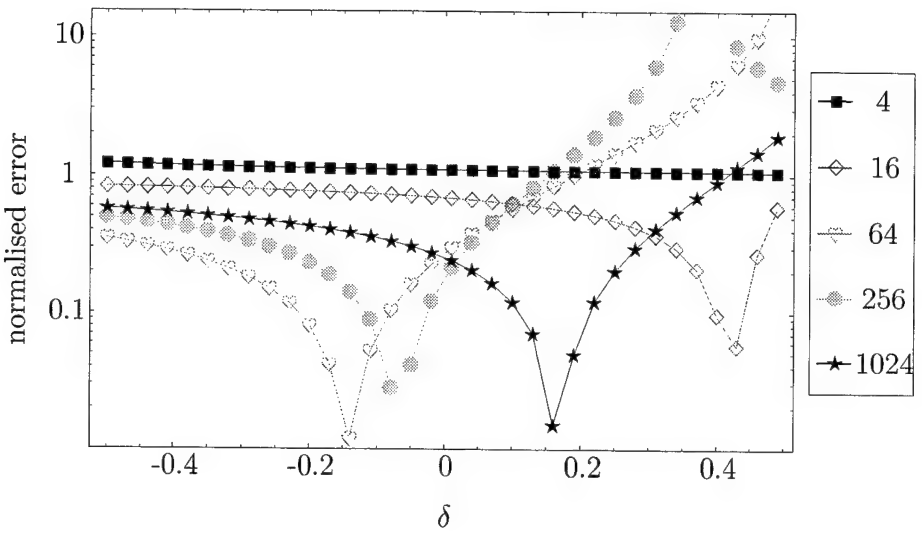


Figure 4.4: Plots of normalised error versus errors in noise estimation $\delta = \delta_u = \delta_v$, for several different approximation orders.

The conditions that guarantee insensitivity to noise estimates (Equations (4.30)–(4.33)) are only valid under the assumption $|\delta_u|, |\delta_u| \ll 1 - \widehat{\phi}\widehat{r}$. This assumption is very restrictive for ill-conditioned systems. In fact it is so restrictive as to throw a shadow of doubt over the corrected LS technique for highly ill-conditioned systems.

After defining the terms bias and consistent, we showed that the LS technique was biased as well as inconsistent, and gave a correction for the inconsistency. In real life we will usually not know the errors in our measurements exactly, for if we did then we would know the exact measurement! As such we have to estimate the noise in the measurements. We showed that the LS correction was sensitive to noise estimations, and developed conditions under which the correction might be sensitive. We saw that bad noise estimates could easily send the correction to infinity.

4.5 LS Approximation with both Input and Output Noise: One-Dimensional System

We briefly look at the special case of a single input, and show that the results developed so far agree with the large amount of literature on this subject.

The exact linear system $z = ax$ is modelled by Ketellapper [16] using the approximate linear system $\widehat{z} = \widehat{a}\widehat{x}$, which is equivalent to the one-dimensional versions of Equations (4.1) and (4.2) respectively. Furthermore, the linear relationship between the measured signal, true signal, and noise, given by Equation (4.3), still holds. Ketellapper assumes normal distributions for true input signal $\widehat{x}_i$, and the input and output noises u_i and w_i, so that the $(\widehat{z}_i, \widehat{x}_i)^T$ for $i = 1, ..., m$ pairs are independent drawings from a bivariate normal distribution. The variance of the input noise, s_{uu}, is assumed to be known.

The least squares (LS) estimator (which Ketellapper terms the "ordinary LS") is given

by $\widehat{a}_{LS} = s_{\widehat{x}\widehat{z}}/s_{\widehat{x}\widehat{x}}$ (this result is well known, see for example Spiegel [32]). Using Equation (4.23) we can write this LS estimate as

$$\widehat{a}_{LS} = \frac{s_{\widehat{x}\widehat{z}}}{s_{xx} + s_{uu}}. \tag{4.34}$$

Ketellapper also states that the corrected LS estimate is given by $\widehat{a}_{CLS} = s_{\widehat{x}\widehat{z}}/(s_{\widehat{x}\widehat{x}} - s_{uu})$, which although less well known has being around for some time [12, p. 170]. Again using Equation (4.23) we can re-write this expression as

$$\widehat{a}_{CLS} = \frac{s_{\widehat{x}\widehat{z}}}{s_{xx}}. \tag{4.35}$$

Setting the correlation to zero ($r = 0$ since we have only one input signal) and substituting $\widehat{a}_{LS}$ in Equation (4.34) for $\widehat{a}$ in Equation (4.21) we obtain Equation (4.35). Thus, the inconsistency correction we developed in § 4.4 is in fact what Ketellapper terms the corrected LS estimate.

If we let η denote the noise to signal ratio s_{uu}/s_{xx}, then Ketellapper shows the inconsistency of the LS technique through the equation

$$\mathcal{E}_{\infty}(\widehat{a}_{LS}) \sim a\left[1 - \frac{\eta}{(1+\eta)}\right], \tag{4.36}$$

where the symbol $\sim$ and the subscripted infinity symbol on the expectation symbolise an asymptotic result. The above expression is a special case of Equation (4.18) with the correlation set to zero ($r = 0$). Furthermore, the variance of the LS estimator (a special case of a result from Schneeweiß [30, Eq. (5.8)]) is given by

$$\mathrm{var}_{\infty}(\widehat{a}_{LS}) \sim \frac{1}{m}\left[\frac{s_{ww}}{s_{xx}}\frac{1}{(1+\eta)} + a^2\frac{\eta}{(1+\eta)^2}\right], \tag{4.37}$$

remember that m denotes the number of points used to develop the LS approximation. Ketellapper also gives the asymptotic variance of the corrected LS estimate as

$$\mathrm{var}_{\infty}(\widehat{a}_{CLS}) \sim \frac{1}{m}\left[\frac{s_{ww}}{s_{xx}}(1+\eta) + a^2\eta(1+2\eta)\right]. \tag{4.38}$$

As we would expect, when the noise in the input signal tends to zero ($\eta \to 0$), the asymptotic variances of the ordinary LS (Equation (4.37)) and corrected LS (Equation (4.38)) converge. Furthermore, the asymptotic bias (inconsistency) of the ordinary LS solution (Equation (4.36)) tends to zero as $\eta \to 0$.

As can be seen from the two asymptotic variances shown above, the ordinary LS estimate (Equation (4.37)) has less variance than the corresponding corrected LS estimate (Equation (4.38)), remember that $\eta \geq 0$. It is for this reason that Ketellapper introduces a measure of "goodness", which he terms the "asymptotic mean square error" (AMSE). The AMSE is defined as the sum of the asymptotic variance and the squared asymptotic bias, that is,

$$\mathrm{AMSE}(\widehat{a}) = \mathrm{var}(\widehat{a}) + [\mathcal{E}(\widehat{a}) - a]^2. \tag{4.39}$$

At first sight the AMSE measure given above may seem questionable, after all it weighs variance and the square of the bias equally. The derivation of the form shown above (see

Equations (5.8)–(5.9)) shows that in fact the mean square error (as Kendall and Stuart [15] term it) is a measure of "closeness" to the true solution.

Ketellapper derives a condition for when to use ordinary LS estimate in preference to the corrected LS estimate. The condition $\mathrm{AMSE}(\widehat{a}_{LS}) < \mathrm{AMSE}(\widehat{a}_{CLS})$ is satisfied if and only if

$$\frac{s_{ww}}{s_{xx}} > a^2\eta\,\frac{(m-4)-\eta(5+2\eta)}{(1+\eta)(2+\eta)}.$$

For m large and η small the above condition becomes $s_{ww}/s_{xx} > a^2\eta\, m/2$. Ketellapper concludes that the strategy to always use corrected LS is appropriate, though not optimal (since under some conditions the ordinary LS procedure may perform better under the AMSE measure).

In this section we have shown the relation between several one-dimensional results found in the literature and the theory that we have developed.

4.6 LS Approximation with both Input and Output Noise: n-Dimensional System

We now generalise the results of the preceding sections to an n-dimensional system, restricting correlation between inputs.

Analogous to the preceding section introduce a linear function of n variables

$$z(x_1, x_2, \ldots, x_n) = \sum_{j=1}^{n} a_j x_j,$$

which we want to approximate using

$$\widehat{z}(\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n) = \sum_{j=1}^{n} \widehat{a}_j \widehat{x}_j.$$

Both the a_j and the $\widehat{a}_j$ are constants. Given a set of m data points $p_i = (\widehat{x}_{i1}, \widehat{x}_{i2}, \ldots, \widehat{x}_{in}; \widehat{z}_i)$ we want to determine the constants $\widehat{a}_j$. The measured input and outputs may be decomposed into true input and output plus input and output noise as follows

$$\widehat{x}_{ij} = x_{ij} + u_{ij} \qquad \text{and} \qquad \widehat{z}_i = z_i + w_i$$

for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, where x_{ij} and z_i are the true inputs and true outputs, and u_{ij} and w_i are their respective noise terms.

For simplicity we assume that the noise terms are uncorrelated with the input, output, and each other (that is, $s_{*u_i} = 0$, where $*$ is any input, output, or noise term except u_i). This assumption is not very restrictive, since if $\epsilon = \max(|s_{*u_i}|)$, then the result from the two-dimensional section still holds, namely $\mathrm{var}(\epsilon) = \mathcal{O}(1/m)$. The least squares (LS) problem is given by the linear system shown below

$$\begin{bmatrix} (1+\eta_1) & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & (1+\eta_2) & \cdots & \gamma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & (1+\eta_n) \end{bmatrix} \begin{bmatrix} \widehat{a}_1 \\ \widehat{a}_2 \\ \vdots \\ \widehat{a}_n \end{bmatrix} = \begin{bmatrix} 1 & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & 1 & \cdots & \gamma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix},$$

where $\eta_i = s_{u_i u_i}/s_{x_i x_i}$ and $\gamma_{ij} = s_{x_i x_j}/s_{x_i x_i}$. The term η_i may be thought of as the ith noise term, and γ_{ij} may be thought of as the normalised covariance between the ith and the jth inputs (normalised against the ith variance). Although the original covariance matrix is symmetric, the normalised covariance matrix is no longer symmetric since we are dividing each row of the covariance matrix by the diagonal term.

The structure of the LS system of equations shown above is more clearly seen if we write

$$(\boldsymbol{I} + \boldsymbol{G} + \boldsymbol{H})\widehat{\boldsymbol{a}} = (\boldsymbol{I} + \boldsymbol{G})\boldsymbol{a}, \tag{4.40}$$

where $\widehat{\boldsymbol{a}} = [\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_n]^T$ and $\boldsymbol{a} = [a_1, a_2, \cdots, a_n]^T$ are vectors. Furthermore, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{H} = \text{diag}(\eta_1, \eta_2, \ldots, \eta_n)$ is the noise matrix (which has elements $h_{ii} = \eta_i$ and $h_{ij} = 0$ for $i \neq j$), and $\boldsymbol{G}$ is the correlation matrix (which has elements $g_{ii} = 0$ and $g_{ij} = \gamma_{ij}$ for $i \neq j$). These three matrices are n-by-n matrices, and $\boldsymbol{H}$ is additionally both diagonal and positive definite. From the above form we clearly see that if $\boldsymbol{H} = \boldsymbol{0}$, then $\boldsymbol{a} = \widehat{\boldsymbol{a}}$.

In their excellent treatise on the total least squares method [35] (see § 5 for more details on this method), Van Huffel and Vandewalle also show that the LS technique is a biased estimate. Consider the system of equations

$$\boldsymbol{X}\boldsymbol{a} = \boldsymbol{z}, \tag{4.41}$$

where the matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ contains the exact input values x_{ij} (that is, measurements without errors) of the n input variables and m samples, the vector $\boldsymbol{a} \in \mathbb{R}^{n \times 1}$ contains the exact coefficients, and the vector $\boldsymbol{z} \in \mathbb{R}^{m \times 1}$ contains the exact output values. The classical LS solution then is given by

$$\widehat{\boldsymbol{a}}_{LS} = \left(\widehat{\boldsymbol{X}}^T \widehat{\boldsymbol{X}}\right)^{-1} \widehat{\boldsymbol{X}}\, \boldsymbol{z}, \tag{4.42}$$

where $\widehat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{U}$, $\widehat{\boldsymbol{a}}$, and $\widehat{\boldsymbol{z}} = \boldsymbol{z} + \boldsymbol{w}$ are the measured values, and hence approximate $\boldsymbol{X}$, $\boldsymbol{a}$, and $\boldsymbol{z}$ respectively.

Van Huffel and Vandewalle (combining the work of two other authors) obtain the following result for convergence of the LS solution [35, p. 232]

$$\widehat{\boldsymbol{a}}_{LS} = \operatorname*{plim}_{m \to \infty} \left[\boldsymbol{I}_n - (\boldsymbol{S}_{\boldsymbol{x}} + \boldsymbol{S}_{\boldsymbol{u}})^{-1} \boldsymbol{S}_{\boldsymbol{u}}\right] \left[\boldsymbol{a} + \boldsymbol{S}_{\boldsymbol{x}}{}^{-1} \boldsymbol{S}_{\boldsymbol{uw}}\right]. \tag{4.43}$$

In the above equation "plim" is the probability limit (that is, converging in probability, see for example [15, p. 3]) and $\boldsymbol{I}_n \in \mathbb{R}^{n \times n}$ is the identity matrix. $\boldsymbol{S}_{\boldsymbol{x}} \in \mathbb{R}^{n \times n}$ is the covariance matrix of the exact inputs, $\boldsymbol{S}_{\boldsymbol{u}} \in \mathbb{R}^{n \times n}$ is the covariance matrix between the input errors, and $\boldsymbol{S}_{\boldsymbol{uw}} \in \mathbb{R}^{n \times 1}$ is the covariance vector between the input errors and the output error. These two matrices and vector are defined respectively by

$$\boldsymbol{S}_{\boldsymbol{x}} = \frac{1}{m}\boldsymbol{X}^T\boldsymbol{X}, \qquad \boldsymbol{S}_{\boldsymbol{u}} = \frac{1}{m}\boldsymbol{U}^T\boldsymbol{U}, \qquad \text{and} \qquad \boldsymbol{S}_{\boldsymbol{uw}} = \frac{1}{m}\boldsymbol{X}^T\boldsymbol{z}. \tag{4.44}$$

Assuming $\boldsymbol{S}_{\boldsymbol{u}} = \sigma^2 \boldsymbol{I}_n$ (that is, the input errors are independent and have the same variance σ^2) and $\boldsymbol{S}_{\boldsymbol{uw}} = 0$ (that is, output and input errors are independent) then Equa-

tion (4.43) simplifies to [35, p. 232]

$$\begin{aligned}\widehat{\boldsymbol{a}}_{LS} &= \left[\boldsymbol{I}_n - \operatorname*{plim}_{m\to\infty} \sigma^2 \left(\boldsymbol{S}_{\boldsymbol{x}} + \sigma^2 \boldsymbol{I}_n\right)^{-1}\right] \boldsymbol{a} \\ &= \left[\boldsymbol{I}_n - \sigma^2 \left(\operatorname*{plim}_{m\to\infty} \boldsymbol{S}_{\widehat{\boldsymbol{x}}}\right)^{-1}\right] \boldsymbol{a},\end{aligned}$$

where $\boldsymbol{S}_{\widehat{\boldsymbol{x}}} \in \mathbb{R}^{n\times n}$ is the covariance matrix of the measured inputs, that is $\boldsymbol{S}_{\widehat{\boldsymbol{x}}} = (1/m)\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}}$. Thus large LS biases result from large errors (either $\boldsymbol{S}_{\boldsymbol{u}}$ or σ large), an ill-conditioned matrix $\boldsymbol{X}$ (that is, inputs almost collinear), or $\boldsymbol{a}$ oriented close to the lowest right singular value of $\boldsymbol{X}$ (that is, $\boldsymbol{a}$ points towards the null space of $\boldsymbol{X}$).

If the covariance matrix $\boldsymbol{S}_{\boldsymbol{u}}$ and covariance vector $\boldsymbol{S}_{\boldsymbol{uw}}$ are known, then the LS bias can be removed using the *corrected LS* technique

$$\widehat{\boldsymbol{a}}_{CLS} = \left(\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}} - m\boldsymbol{S}_{\boldsymbol{u}}\right)^{-1}\left(\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{z}} - m\boldsymbol{S}_{\boldsymbol{uw}}\right). \tag{4.45}$$

Assuming input and output noise are uncorrelated (that is, $\boldsymbol{S}_{\boldsymbol{uw}} = \boldsymbol{0}$) then the two-dimensional version of the above equation reduces to Equation (4.25).

Van Huffel and Vandewalle show that the under the assumption of independently and identically distributed input and output errors (that is, $\boldsymbol{S}_{\boldsymbol{u}} = \sigma^2\boldsymbol{I}_n$ and $\operatorname{var}(\boldsymbol{z}) = \sigma$), the corrected LS and TLS solutions asymptotically yield the same consistent estimate $\widehat{\boldsymbol{a}}$ of the exact solution $\boldsymbol{a}$. Naidu [22] independently shows that the corrected LS (or *alternative LS* as Naidu terms it) is consistent.

Compare the LS solution given by Equations (4.42) with the corrected LS solution given by Equation (4.45). The above expression may be thought of as a "method-of-moments" estimator [35] since $\mathcal{E}(\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}} - m\boldsymbol{S}_{\boldsymbol{u}}) = \boldsymbol{X}^T\boldsymbol{X}$ and $\mathcal{E}(\widehat{\boldsymbol{X}}\,\widehat{\boldsymbol{z}} - m\boldsymbol{S}_{\boldsymbol{uw}}) = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a}$, where $\mathcal{E}(\cdot)$ is the expectation (or mean). Johnston [12, Eq. (6-47)] gives the same result for the corrected LS, and goes on to state that the intercept term $\widehat{\boldsymbol{a}}_0$ can be estimated by passing the solution plane through the sample means, or the centroid as Nievergelt [23] terms it. (For further details on this statement about the centroid see § 5, especially Figure 5.1.)

We now look at special cases of the LS system (4.40).

4.6.1 Each Input Correlated with at Most One Other Input

To make the problem more tractable, we assume that all inputs are correlated with at most one other input. For example, if $\gamma_{ij}, \gamma_{ji} \neq 0$, then $\gamma_{ik} = \gamma_{ki} = 0$ (for $k \neq j$) and $\gamma_{kj} = \gamma_{jk} = 0$ (for $k \neq i$), that is, all γ's on the ith row and columns, and jth row and column of the matrix $\boldsymbol{G}$ are zero except for γ_{ij} and γ_{ji}. This restriction implies that we have at most n (for n even) or $n-1$ (for n odd) input correlations. Below is an example

of this restriction for a 7-by-7 matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \gamma_{16} & 0 \\ 0 & 0 & 0 & 0 & \gamma_{25} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \gamma_{37} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \gamma_{52} & 0 & 0 & 0 & 0 & 0 \\ \gamma_{61} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma_{73} & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{4.46}$$

Note that re-arranging the rows of the above system yields a diagonal matrix. We now see that the condition of "correlation to at most one other input" is satisfied whenever the normalised covariance matrix is diagonisable.

If we further assume that all noise terms η have the same magnitude (that is, $\eta_i = \eta$), then the general solutions are

$$\widehat{a}_i = a_i \left(1 - \eta \frac{(1 + \eta - \frac{a_j}{a_i}\gamma_{ij})}{(1 + \eta - \gamma_{ij})(1 + \eta + \gamma_{ij})} \right) \tag{4.47}$$

and

$$\widehat{a}_j = a_j \left(1 - \eta \frac{(1 + \eta - \frac{a_i}{a_j}\gamma_{ji})}{(1 + \eta - \gamma_{ji})(1 + \eta + \gamma_{ji})} \right) \tag{4.48}$$

if $\gamma_{ik} = \gamma_{ki} = 0$ (for $k \neq j$) and $\gamma_{jk} = \gamma_{kj} = 0$ (for $k \neq i$). Notice that Equation (4.47) has the same form as Equation (4.18).

On the other hand, if $\gamma_{ik} = \gamma_{ki} = 0$ for $k = 1, 2, \ldots, n$, then the above expressions simplify to

$$\widehat{a}_i = \frac{a_i}{1 + \eta}. \tag{4.49}$$

Using the example correlation matrix given by Equation (4.46) and Equations (4.47)–(4.49), the coefficients of the least squares (LS) approximation $\widehat{a}_i$ are given by

$$\begin{bmatrix} \widehat{a}_1/a_1 \\ \widehat{a}_2/a_2 \\ \widehat{a}_3/a_3 \\ \widehat{a}_4/a_4 \\ \widehat{a}_1/a_5 \\ \widehat{a}_1/a_6 \\ \widehat{a}_1/a_7 \end{bmatrix} = \begin{bmatrix} 1 - \eta(1 + \eta - \frac{a_6}{a_1}\gamma_{16}) / \left[(1 + \eta - \gamma_{16})(1 + \eta + \gamma_{16})\right] \\ 1 - \eta(1 + \eta - \frac{a_5}{a_2}\gamma_{25}) / \left[(1 + \eta - \gamma_{25})(1 + \eta + \gamma_{25})\right] \\ 1 - \eta(1 + \eta - \frac{a_7}{a_3}\gamma_{37}) / \left[(1 + \eta - \gamma_{37})(1 + \eta + \gamma_{37})\right] \\ a_4/(1 + \eta) \\ 1 - \eta(1 + \eta - \frac{a_2}{a_5}\gamma_{52}) / \left[(1 + \eta - \gamma_{52})(1 + \eta + \gamma_{52})\right] \\ 1 - \eta(1 + \eta - \frac{a_1}{a_6}\gamma_{61}) / \left[(1 + \eta - \gamma_{61})(1 + \eta + \gamma_{61})\right] \\ 1 - \eta(1 + \eta - \frac{a_3}{a_7}\gamma_{73}) / \left[(1 + \eta - \gamma_{73})(1 + \eta + \gamma_{73})\right] \end{bmatrix}$$

If all the η_i do not have the same magnitude, then the more general forms of the two-dimensional solutions given by Equations (4.16) and (4.17) may be substituted for Equations (4.47) and (4.48) respectively (remembering to make the appropriate change in notation).

4.6.2 Every Input is Correlated to All Other Inputs

The approximation we make in this section is not as useful as the results of the previous section, since we assume that all correlations are of the same magnitude (quite a severe restriction).

Assume that every input is correlated to all other inputs by exactly the same amount, that is, $\gamma_{ij} = \gamma$. Furthermore, assume that all noise terms have the same magnitude, that is, $\eta_i = \eta$. Under these two assumptions, Equation (4.40) reduces to

$$(\boldsymbol{I} + \gamma \boldsymbol{Z} + \eta \boldsymbol{I})\widehat{\boldsymbol{a}} = (\boldsymbol{I} + \gamma \boldsymbol{Z})\boldsymbol{a},$$

where $\boldsymbol{Z}$ is an n-by-n matrix that is filled with unity except for the diagonal terms which contain zeroes, that is, the elements of $\boldsymbol{Z}$ are $z_{ij} = 1 - \delta_{ij}$ (where $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$). The least squares (LS) coefficients $\widehat{a}_i$ can be shown to have the solution

$$\widehat{a}_i = a_i \left\{ 1 - \eta \frac{\left[1 + \eta - \gamma\left(1 - n + \sum_{j=1}^{n} \frac{a_i}{a_j}\right)\right]}{(1 + \eta - \gamma)\left[1 + \eta + (n-1)\gamma\right]} \right\} \qquad \text{for } i = 1, 2, \ldots, n.$$

We began this section by generalising the two-dimensional results to n-dimensions. The results obtained earlier for the two-dimensional case were found to have a natural extension to higher dimensions. In particular, the LS and the corrected LS procedures were seen to be, respectively, inconsistent and consistent. We also developed solutions for special cases of the correlation matrix.

4.7 Effects of Noise on the Matrix Technique

In this section we show why the matrix technique out-performs the vector technique for noisy systems with correlated inputs.

Consider the two-input (three strain gauge) system shown in Figure 3.1 (on page 10). We change the notation of the coefficient in this section, so that the *true* and *approximate* stresses are related to the external loads by

True		*Approximate*	
$\sigma_1 = A_1 P_x + B_1 P_y,$		$\widehat{\sigma}_1 = \widehat{a}_1 \widehat{P}_x + \widehat{b}_1 \widehat{P}_y,$	(4.50)
$\sigma_2 = A_2 P_x + B_2 P_y,$		$\widehat{\sigma}_2 = \widehat{a}_2 \widehat{P}_x + \widehat{b}_2 \widehat{P}_y,$	(4.51)
$\sigma_0 = A_0 P_x + B_0 P_y,$	and	$\widehat{\sigma}_0 = \widehat{a}_0 \widehat{P}_x + \widehat{b}_0 \widehat{P}_y.$	(4.52)

The true solutions are shown on the left and the approximate solutions are shown on the right.

Solving for the true loads P_x and P_y in terms of the stresses σ_1 and σ_2 using Equations (4.50) and (4.51), then substituting this result into σ_0 in Equation (4.52) gives

$$\sigma_0 = c_1 \sigma_1 + c_2 \sigma_2,$$

where

$$c_1 = \frac{A_0B_2 - A_2B_0}{A_1B_2 - A_2B_1} \qquad \text{and} \qquad c_2 = -\frac{A_0B_1 - A_1B_0}{A_1B_2 - A_2B_1}. \tag{4.53}$$

(Note these same expressions were derived earlier using a different notation for the coefficients, see Equations (3.1) and (3.2).) Similarly for the approximate system we have that

$$\sigma_0 = \widehat{c}_1\sigma_1 + \widehat{c}_2\sigma_2,$$

where

$$\widehat{c}_1 = \frac{\widehat{a}_0\widehat{b}_2 - \widehat{a}_2\widehat{b}_0}{\widehat{a}_1\widehat{b}_2 - \widehat{a}_2\widehat{b}_1} \qquad \text{and} \qquad \widehat{c}_2 = -\frac{\widehat{a}_0\widehat{b}_1 - \widehat{a}_1\widehat{b}_0}{\widehat{a}_1\widehat{b}_2 - \widehat{a}_2\widehat{b}_1}. \tag{4.54}$$

From the least squares (LS) section, Equations (4.9) and (4.10), we know the solution for a noisy two-input/single-output system is given by

$$\sigma_* = \widehat{a}_*P_x + \widehat{b}_*P_y,$$

where

$$\widehat{a}_* = \frac{A_*(s_{xx}s_{yy} - s_{xy}^2 + s_{vv}s_{xx}) + B_*s_{vv}s_{xy}}{(s_{uu} + s_{xx})(s_{vv} + s_{yy}) - s_{xy}^2} + \mathcal{O}(\epsilon)$$

and

$$\widehat{b}_* = \frac{B_*(s_{xx}s_{yy} - s_{xy}^2 + s_{uu}s_{yy}) + A_*s_{uu}s_{xy}}{(s_{uu} + s_{xx})(s_{vv} + s_{yy}) - s_{xy}^2} + \mathcal{O}(\epsilon).$$

Introducing the notation

$$s_1 = s_{xx}s_{yy} - s_{xy}^2 + s_{vv}s_{xx}, \qquad s_2 = s_{vv}s_{xy}, \qquad s_3 = (s_{uu} + s_{xx})(s_{vv} + s_{yy}) - s_{xy}^2,$$
$$s_4 = s_{xx}s_{yy} - s_{xy}^2 + s_{uu}s_{yy}, \qquad \text{and} \qquad s_5 = s_{uu}s_{xy},$$

we can then express the coefficients $\widehat{a}_*$ and $\widehat{b}_*$ more concisely as

$$\widehat{a}_* = \frac{A_*s_1 + B_*s_2}{s_3} + \mathcal{O}(\epsilon) \qquad \text{and} \qquad \widehat{b}_* = \frac{B_*s_4 + A_*s_5}{s_3} + \mathcal{O}(\epsilon).$$

Substituting the above expressions into Equation (4.54) gives the numerator and denominator of $\widehat{c}_1$ respectively as

$$\widehat{a}_0\widehat{b}_2 - \widehat{a}_2\widehat{b}_0 = (A_0B_2 - A_2B_0)\frac{s_1s_4 - s_2s_5}{s_3^2} + \mathcal{O}(\epsilon)$$

and

$$\widehat{a}_1\widehat{b}_2 - \widehat{a}_2\widehat{b}_1 = (A_1B_2 - A_2B_1)\frac{s_1s_4 - s_2s_5}{s_3^2} + \mathcal{O}(\epsilon).$$

Hence the coefficient $\widehat{c}_1$ has the form

$$\widehat{c}_1 = \frac{A_0B_2 - A_2B_0}{A_1B_2 - A_2B_1} + \mathcal{O}(\epsilon),$$

which reduces to the form

$$\widehat{c}_1 = c_1 + \mathcal{O}(\epsilon)$$

if Equation (4.53) is used. Similarly the coefficient $\widehat{c}_2$ has the form $\widehat{c}_2 = c_2 + \mathcal{O}(\epsilon)$. In other words, the matrix technique is not noise sensitive (to order ϵ) when the inputs are correlated. We have eliminated the troublesome denominator s_3 using additional information.

The above analysis explains the error plots shown in Figure 3.7. As can be seen the vector technique is an inconsistent estimator of the true solution, whereas the matrix technique is a consistent estimator. Furthermore, the slope of the approximation order versus the normalised error is -1/2 as predicted from the order ϵ analysis in § 4.2 (see Equation (4.11)).

In the above analysis we have implicitly assumed that the denominator of $\widehat{a}_*$ and $\widehat{b}_*$ is non-zero, that is $s_3 = (s_{uu} + s_{xx})(s_{vv} + s_{yy}) - s_{xy}^2 \neq 0$. This assumption is true for any real system, since $s_{uu}, s_{vv} > 0$, that is, both inputs have noise. However, $s_3 \neq 0$ is satisfied even for ideal noiseless systems ($s_{uu} = s_{vv} = 0$) provided the two inputs are not totally correlated, that is, provided $P_x \neq P_y$ for at least one point.

In this section the superior performance of the matrix technique was shown to be due to the elimination of the troublesome denominator (a function of correlation), making the matrix technique consistent.

4.8 Relation between Condition Number and Output Correlation

In this section we investigate the relationship between the condition number of the underlying linear system and the correlation of the measured stresses (or outputs). We will see that an ill-conditioned system does not necessarily imply output correlation. Neither is the reverse statement true, that is, measured stress (or output) correlation does not necessarily imply an ill-conditioned system.

We begin by determining the condition number of the simple five member truss we investigated in § 3.4. We numerically simulated the measurement of stress of that truss, and using these "measured" stresses (true stress plus noise) in members OA and AB we predicted the stress in member OC. From Equations (3.3) and (3.5) we have that

$$\boldsymbol{\sigma} = \boldsymbol{A}\boldsymbol{p} \tag{4.55}$$

where $\boldsymbol{\sigma} = [\sigma_{OA}, \sigma_{AB}]^T$, $\boldsymbol{p} = [P_x, P_y]^T$, and

$$\boldsymbol{A} = \begin{bmatrix} \left(1 - \frac{\gamma}{\beta}\right) & -\gamma\left(1 - \frac{\alpha}{\beta}\right) \\ \left(1 - \frac{\gamma}{\beta}\right) & -\alpha\left(1 - \frac{\gamma}{\beta}\right) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{4.56}$$

Using the results from § 2.3 we can determine the condition number of the above system. We know that for an infinite condition number (of a two-by-two system) we require that

$|\boldsymbol{A}| = 0$, which yields the two conditions $\alpha = \gamma$ and $\beta = \gamma$. These two conditions were shown as special cases of the truss (see Figure 3.5, which shows either member AC or BC as vertical).

Let's now consider the correlation between the two outputs (σ_{OA} and σ_{AB}). From Equation (4.15) the correlation coefficient between two variables x and y is defined as $r(x,y) = s_{xy}/\sqrt{s_{xx}s_{yy}}$, where $s_{xy} = (1/n)\sum_{i=1}^{n} x_i y_i$ is the covariance between x and y for n sample points. For conciseness define the following notation: the covariance between external loads $s_{xx} = \mathrm{cov}(P_x, P_x)$, $s_{yy} = \mathrm{cov}(P_y, P_y)$, and $s_{xy} = \mathrm{cov}(P_x, P_y)$; the covariance between stresses $s_{11} = \mathrm{cov}(\sigma_{OA}, \sigma_{OA})$, $s_{22} = \mathrm{cov}(\sigma_{AB}, \sigma_{AB})$, and $s_{12} = \mathrm{cov}(\sigma_{OA}, \sigma_{AB})$; and the correlation between the two outputs $r_\sigma = r(\sigma_{OA}, \sigma_{AB})$. Using the definition of covariance we then have that

$$\begin{aligned} s_{11} &= \frac{1}{n}\sum_{i=1}^{n} (a_{11}P_x + a_{12}P_y)^2 \\ &= a_{11}^2 s_{xx} + 2a_{11}a_{12}s_{xy} + a_{12}^2 s_{yy}, && (4.57) \\ s_{22} &= a_{21}^2 s_{xx} + 2a_{21}a_{22}s_{xy} + a_{22}^2 s_{yy}, \qquad \text{and} && (4.58) \\ s_{12} &= a_{11}a_{21}s_{xx} + (a_{11}a_{22} + a_{12}a_{21})\, s_{xy} + a_{12}a_{22}s_{yy}, && (4.59) \end{aligned}$$

where a_{11}, a_{12}, a_{21}, and a_{22} are the coefficients of the matrix $\boldsymbol{A}$ given in Equation (4.56). We may then write down the correlation coefficient between the two outputs σ_{OA} and σ_{AB} as

$$r_\sigma^2 = \frac{\left\{\beta\left[s_{xx} - (\alpha+\gamma)s_{xy} + \alpha\gamma s_{yy}\right] - \gamma\left[s_{xx} - 2\alpha s_{xy} + \alpha^2 s_{yy}\right]\right\}^2}{\left(s_{xx} - 2\alpha s_{xy} + \alpha^2 s_{yy}\right)\left\{(\beta-\gamma)\left[(\beta-\gamma)s_{xx} + 2\gamma(\alpha-\beta)s_{xy}\right] + (\alpha-\beta)^2\gamma^2 s_{yy}\right\}}. \tag{4.60}$$

We want to determine when the two outputs σ_{OA} and σ_{AB} are correlated (that is, $r_\sigma = 1$). Using the above equation, $r_\sigma^2 = 1$ is satisfied whenever the condition

$$\beta^2(\alpha-\gamma)^2(s_{xy}^2 - s_{xx}s_{yy}) = 0$$

is met. Thus there are three ways the outputs σ_{OA} and σ_{AB} can be completely correlated; if $\beta = 0$ or $\alpha = \gamma$ (two of the three conditions for ill-conditioning), or if the two forcing loads P_x and P_y are completely correlated (that is, $s_{xy}^2 - s_{xx}s_{yy} = 0$). However, $\beta = 0$ cannot be satisfied under the conditions imposed earlier ($|\beta| > 0$, see page 15).

We now see that ill-conditioning does not imply output correlation for the truss system developed in § 3.3. For example, if $\beta = \gamma$ then $\kappa = \infty$ but $r_\sigma \neq 1$. Furthermore, output correlation does not imply system ill-conditioning, for example, if $P_x \propto P_y$ then $r_\sigma = 1$ but $\kappa \neq \infty$. These results, however, may be due to the fact that the truss system developed in § 3.3 is (coincidentally) a special case, as we now show.

Let's consider the same system given by Equation (4.55), except the matrix $\boldsymbol{A}$ is now given by a general two-by-two matrix. By definition the square of the correlation coefficient is $r_\sigma^2 = s_{12}^2/(s_{11}s_{22})$, using s_{11}, s_{22}, and s_{12} as defined by Equations (4.57)–(4.59) we obtain

$$r_\sigma^2 = \frac{\left[a_{11}a_{21}s_{xx} + (a_{11}a_{22} + a_{12}a_{21})\, s_{xy} + a_{12}a_{22}s_{yy}\right]^2}{\left(a_{11}^2 s_{xx} + 2a_{11}a_{12}s_{xy} + a_{12}^2 s_{yy}\right)\left(a_{21}^2 s_{xx} + 2a_{21}a_{22}s_{xy} + a_{22}^2 s_{yy}\right)}.$$

Solving for the case when the two outputs σ_{OA} and σ_{AB} are completely correlated (that is, $r_\sigma^2 = 1$) gives the condition

$$\begin{aligned}(a_{11}a_{22} - a_{12}a_{21})^2 (s_{xy}^2 - s_{xx}s_{yy}) &= 0 \\ |\boldsymbol{A}|^2(s_{xy}^2 - s_{xx}s_{yy}) &= 0.\end{aligned}$$

Thus it appears as if the truss system we developed in § 3.3 is a special case, and in general ill-conditioning implies correlation, but not the reverse. (The other possibility is that all trusses lead to the special case shown above. This idea was not investigated further.)

The above analysis was undertaken for the limiting case of completely correlated measured outputs ($r_\sigma = 1$). We need to ask the same questions of partially correlated outputs.

4.8.1 Partial Correlation: Simulation of Random Truss Geometries

Analytically solving for conditions such as when κ approaches infinity and r_σ nears unity would result in a complex set of conditions. Instead a simulation was conducted to determine the effects of configuration (of the five-member truss developed in § 3.3) on both the output correlation and the condition number.

Before we look at the relation between output correlation and condition number, let's transform the range of the condition number, $\kappa \in [1, \infty)$, to that of output correlation, $r_\sigma \in [0, 1]$. One transformation that will achieve this mapping is given by

$$\kappa' = 1 - \frac{1}{\kappa^2}, \tag{4.61}$$

which we term the *transformed condition number*. (The only reason κ squared was used in the above equation was to eliminate the square root in the expression for condition number given by Equation (2.18).)

We begin by evaluating the form of the transformed condition number for the truss. From Equations (2.18) and (4.56) we have that the condition number for the five-member truss system developed in § 3.3 is given by

$$\kappa = \sqrt{1 + \frac{2}{\widehat{\kappa} - 1}}, \tag{4.62}$$

where

$$\widehat{\kappa}^2 = \frac{\left\{2(\alpha^2+1)\gamma^2 + \beta^2(\alpha^2+\gamma^2+2) - 2\beta\gamma\left[\alpha(\alpha+\gamma)+2\right]\right\}^2}{4(\beta-\gamma)^2\left[\alpha(\beta-2\gamma)+\beta\gamma\right]^2 + \left\{2(\alpha^2-1)\gamma^2 + \beta^2(\alpha^2+\gamma^2-2) - 2\beta\gamma\left[\alpha(\alpha+\gamma)-2\right]\right\}^2}.$$

Using Equations (4.61) and (4.62) we then have that the transformed condition number is given by

$$\kappa' = \frac{2}{1+\widehat{\kappa}}.$$

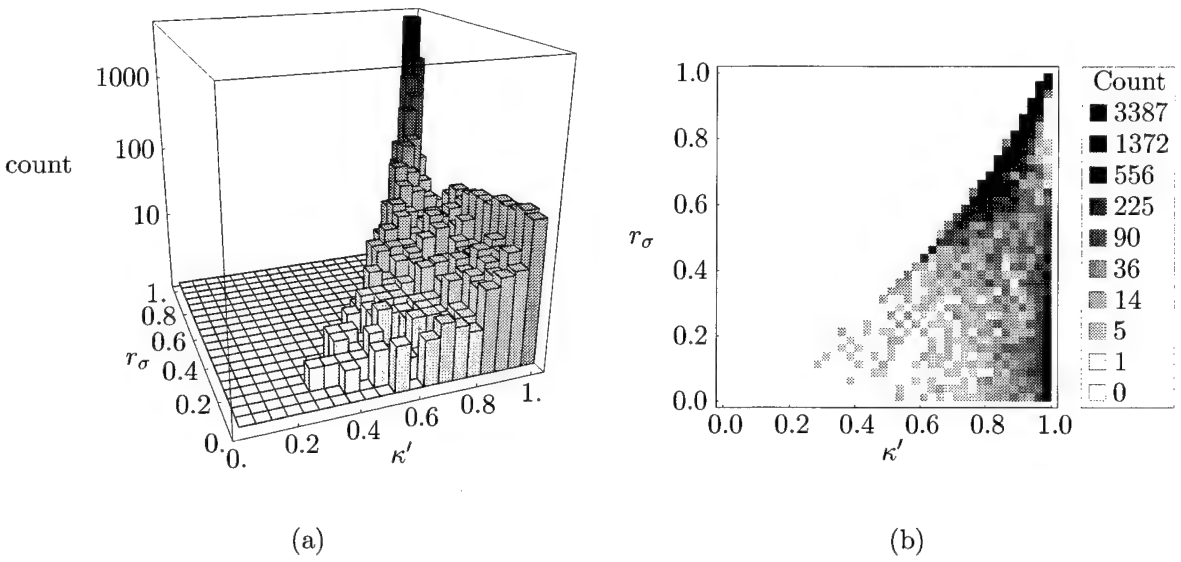
We now derive the form of the output correlation. If we assume the two inputs (P_x and P_y) are completely uncorrelated (that is, $s_{xy} = 0$), then Equation (4.60) simplifies to

$$r_\sigma^2\Big|_{s_{xy}=0} = \frac{\left[(\beta-\gamma)s' - \alpha(\alpha-\beta)\gamma\right]^2}{(s'+\alpha^2)\left[(\beta-\gamma)^2 s' + (\alpha-\beta)^2\gamma^2\right]},$$

where s' is the ratio of the two input covariances, that is, $s' = s_{xx}/s_{yy}$.

Now that we have both the transformed condition number and the output correlation we can compare how these two quantities are related using a simulation of random truss geometry.

For our simulation the truss' configuration was varied by randomly choosing values for the parameters α, β, and γ (these parameters uniquely define the truss' non-dimensional geometry, see § 3.3). The parameters α, β, and γ were randomly selected using a uniform distribution with the following ranges, $\gamma \in [\epsilon, 1]$, $\beta \in [\epsilon, 1-\epsilon]$, and $\alpha \in [\epsilon, \beta]$, with ϵ set to $\epsilon = 0.01$ (an arbitrary choice of a small number). Using these random truss parameters 10 000 condition-correlation coordinate pairs (κ', r_σ) were generated, we term these coordinate pairs "points". Finally, in the evaluation of the output correlation r_σ we have selected the input forces P_x and P_y to be completely uncorrelated (that is, $s_{xy} = 0$), and the ratio of the input variances to be unity (that is, $s' = s_{xx}/s_{yy} = 1$).



(a) (b)

Figure 4.5: Bin counts of 10 000 *coordinate pairs* (κ', r_σ) *determined using random values of* α, β, *and* γ.

Figure 4.5 shows the result of bin counts on this 10 000 point set. Both plots shows how many of the 10 000 points occurred in any one κ'-r_σ bin, and the shading represents the number of points in any one bin (the darker the shade the more points). The transformed condition number κ' and output correlation r_σ ranges were partitioned into 20 equal intervals for the three-dimensional plot and 40 equal intervals for the two-dimensional plot. Due to this difference in partitioning, the shading in the three- and two- dimensional plots represent different point counts. Finally, note that the vertical scale on the three-dimensional plot and the shading in both plots is logarithmic.

We see from Figure 4.5 that the majority of geometric configurations result in systems that tend to be ill-conditioned ($\kappa' \approx 1$) and result in correlated outputs ($r_\sigma \approx 1$). This

result may be due to the fact that the vast majority of random geometric configurations would have a skewed appearance, and hence we would expect the resulting truss system to be ill-conditioned. If the range of the geometric parameters (α, β, and γ) are changed, then the resulting bin count differs to that shown in Figure 4.5. However, several features remain constant even when the geometric parameter ranges are changed.

- The correlation between the two outputs, r_σ, appears to always be less than the transformed condition number κ'. (There is no shading above the $\kappa' = r_\sigma$ line.)
- Most geometric configurations have a higher probability of tending towards ill-conditioning and high output-correlation than any other κ'-r_σ combination. (The dark shading is concentrated around the top righthand corner.)
- The degree of ill-conditioning κ' and output correlation r_σ appear to be highly correlated. (Notice the black shading near the $\kappa' = r_\sigma$ line.)

Figure 4.5 also shows us that even if the truss geometry is ill-conditioned (that is, $\kappa' \approx 1$), this doesn't necessarily imply that the outputs will be correlated (that is, it is possible to obtain $r_\sigma \ll 1$). This observation is highlighted by the fact that the shading for $\kappa' \approx 1$ occurs around both r_σ small and $r_\sigma \ll 1$. Additionally, these preliminary simulations suggest that the output measurements will never be more correlated than the ill-conditioning of the underlying truss configuration (that is, $r_\sigma < \kappa'$).

In this section we determined both the correlation and condition number of the five member truss used for simulations. We showed that correlated input signals do not imply an ill-conditioned system, and conversely, that an ill-conditioned system does not imply correlated input signals. Due to the complexity in developing conditions for near unit correlation and near ill-conditioning, we resorted to simulations. As expected, we found (using a simulation of thousands of random truss geometries) that correlation and condition number were loosely related. Additionally, and unexpectedly, the simulations suggest that the input signal correlation is always better than the ill-conditioning of the system, which is good news.

4.9 Simulation Results of the Corrected-Vector Technique

In this subsection we implement the result of this section, namely the corrected least squares (CLS) procedure applied to the ordinary-vector technique. (In this subsection we rename the "vector" technique from previous sections as the "ordinary-vector" technique to clearly distinguish it from the corrected-vector technique.)

A simulation of random load inputs into the five member truss (developed in § 3.3) was undertaken. The simulation was the same as that carried out earlier in § 3.4. Different *approximation orders* (the number of points in the LS solution) were tested for accuracy, and within each approximation order 1000 solution batches were obtained. From these 1000 batch sets, the distribution of errors could be ascertained by determining the outliers and quartiles of the sets. (See § 3.4 for further details.)

The median solution in this subsection was calculated using the L_1 (or spatial) median, where the distance from the median to the 1000 solutions was minimised. (For more details on the L_1 median see Small [31] or for a brief account Martin [21].) In contrast, in § 3.4 the median solution was calculated as the median of each component. Let $\boldsymbol{x}$ denote an n-dimensional vector having components $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{x}_i$ be one of the m data points used to determine the median. Then the L_1 median and component median are defined respectively as

$$\min_{\boldsymbol{x}} \sum_{i=1}^{m} ||\boldsymbol{x} - \boldsymbol{x}_i||_2, \qquad \text{and} \qquad (\bar{x}_{i1}, \bar{x}_{i2}, \ldots, \bar{x}_{in}),$$

where x_{ij} is the jth component of the ith data point and $\bar{x}_{i2} = \text{median}_i(x_{i2})$, for example, is the median of the data's second component (over all the data points $i = 1, \ldots, m$).

Figures 4.6 illustrates sample solutions for the ordinary-vector, matrix, and corrected-vector technique. These solutions are plotted as the a_1 and a_2 coefficients of Equation (3.8). Within each of these plots the effect of approximation order (the number of points used to develop the solution) is depicted. Six solutions were obtained for each approximation order of 2, 8, 32, 128, or 512 points with associated star, square, circle, triangle, and cross symbols respectively.

The most striking feature of these illustrations is that the solutions appear to lie on (or close to) a straight line. A brief experimentation with system condition number seems to suggest that this phenomenon is a product of an ill-conditioned system. (Perhaps this straight line aligns with the null-space of the system?)

Due to the large scale changes, the three plots shown in Figure 4.6 have been plotted on a log-like scale. The spatial median of the highest order approximation (512 points) for each technique was set as the origin of the inverse hyperbolic sine function ($\sinh^{-1}$). Mathematically, if $\boldsymbol{x}$ represents the original space, then the transformed space $\boldsymbol{x}'$ is given by

$$\boldsymbol{x}' = \sinh^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_i),$$

where $\bar{\boldsymbol{x}}_i = \text{median}_i(\boldsymbol{x}_i)$ is the spatial median of the data points $\boldsymbol{x}_i$. This hyperbolic mapping maintains a near-linear spacing around the origin (the median in our case) and produces a log-like expansion away from the origin. Unfortunately, compared to the linear plots these log-like plots de-emphasise the difference in solution spread as the approximation order changes. However, this hyperbolic transformation is required to show the different solutions, with a reasonable spread, on the one plot.

The centre of the horizontal and vertical straight lines (the cross hair) denotes the location of the exact solution (given by the coefficients of Equation (3.9)).

We again see the characteristics described earlier (in § 3.4, especially Figure 3.7) for the ordinary-vector and matrix techniques. The spread of ordinary-vector solutions is small, and decreases with increasing approximation order. However, the ordinary-vector solution moves away from the exact solution as the approximation order increases. In contrast, the matrix technique has a larger spread, but its solutions contract around the exact solution as the approximation order increases.

The corrected-vector technique is similar to the matrix technique in that its solutions have a large spread and they contract around the exact solution as the approximation order

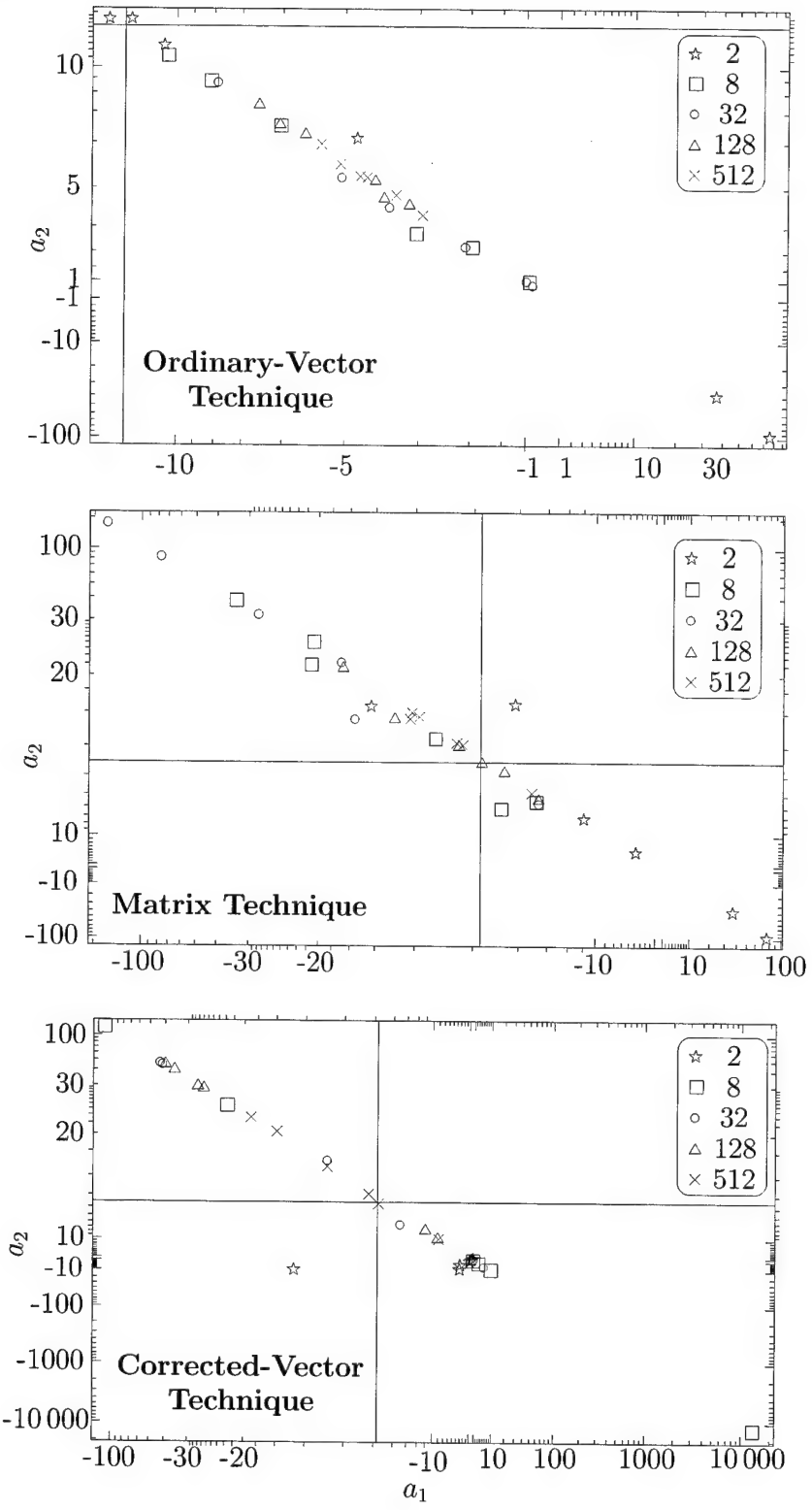


Figure 4.6: Sample solutions (a_1 and a_2 coefficients) for different approximation orders (number of points used to develop the solution). The cross-hairs denote the location of the exact solution.

increases. The corrected-vector technique, however, has a far larger spread (by two orders of magnitude) than the matrix technique. Additionally, the corrected-vector technique only begins to contract around the exact solution once the approximation order is large.

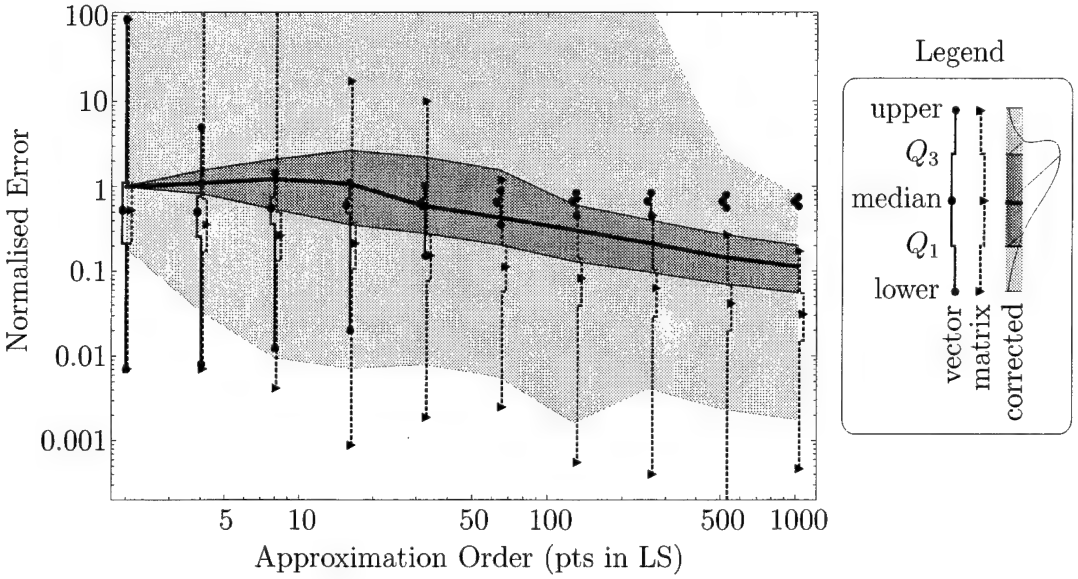


Figure 4.7: Comparison of the normalised error of the corrected-vector technique (shaded planes) to the ordinary-vector technique (circles) and matrix technique (triangles). The distribution of errors is illustrated by the error outliers (denoted as lower and upper), the first and third quartiles (Q_1 and Q_3), and the median.

The distribution of the normalised error for the corrected-vector technique is plotted in Figure 4.7 (for a comparison see Figure 3.7 on page 18). For each of the approximation orders (2, 4, 8, ..., 1024 points) we plot the smallest outlier (denote as "lower"), the first quartile (Q_1), the median, the third quartile (Q_3), and the largest outlier (denoted as "upper"), which allows us to determine the distribution of errors.

The distribution of errors for the corrected-vector technique is shown as shaded planes. In contrast, the distribution of errors for the ordinary-vector technique and matrix technique use box-plots. In these box-plots the symbols (either circles or triangles for the ordinary-vector and matrix techniques respectively) are placed at the outliers and the median. The first and third quartiles are located at the lower and upper extremes of the box.

As can be seen in Figure 4.7 the corrected-vector technique is not as accurate as the matrix technique. In fact, for low order approximations the ordinary-vector technique also produces smaller errors than the corrected-vector technique. In general we see that for high order approximations the errors of the corrected-vector technique are roughly an order of magnitude larger than the matrix technique. Finally, note the large spread of errors exhibited by the corrected-vector technique

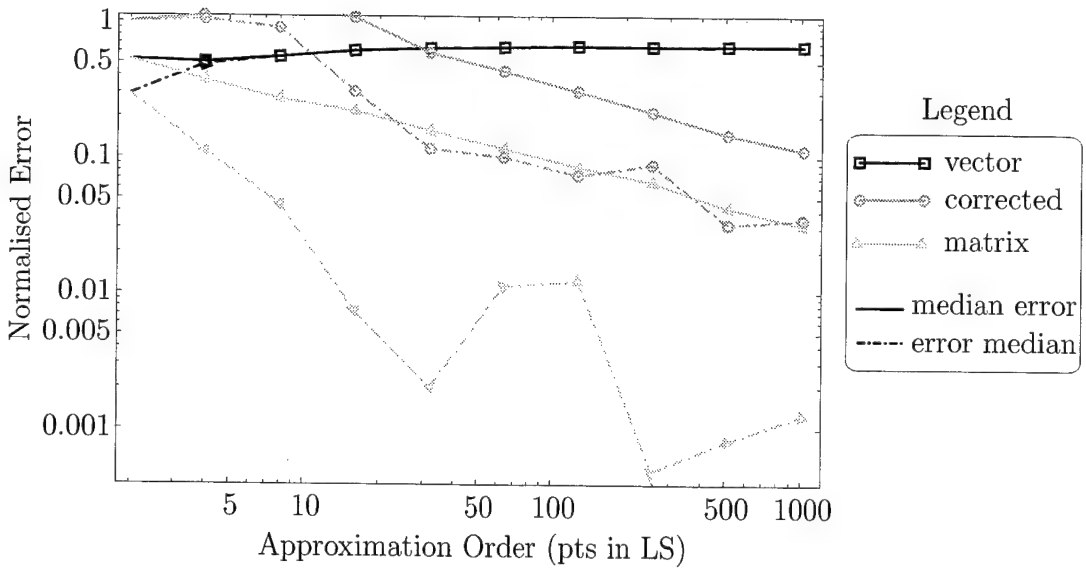


Figure 4.8: Comparison of the normalised error of the median solution (dashed lines) for the corrected-vector (triangles), ordinary-vector (circles), and matrix (squares) techniques. For comparison the median of the normalised errors (solid lines) are also shown.

Figure 4.8 shows how both the median of the errors and the error of the median solution vary with approximation order (see page 19 for a discussion on the difference between these two errors). We see that both the corrected-vector and the matrix techniques' error improves by taking the median of the 1000 solutions evaluated for each approximation order. For these two techniques the error of the median improves by an order of magnitude (approximately) when compared to the median of the errors.

We have seen that although the corrected-vector technique is more accurate than the ordinary-vector technique (for large approximation orders), the accuracy is still inferior to that of the matrix technique. This inferior accuracy is most likely due to the sensitivity of the correction to errors in noise estimations (as outlined earlier).

4.10 Simulation Results of the Surrogate-Matrix Technique

Although the matrix technique yields good results, it requires the use of external loads during calibration. We saw earlier in § 4.7 that the reason the matrix technique produced good results was that the troublesome denominator was eliminated using external loads. A question then naturally arises: From the perspective of the linear system what is so special about the external loads? The simple answer is nothing. Why then don't we use additional information from elsewhere in the linear system (for example, extra gauges instead of external loads) to eliminate the troublesome denominator. In the current subsection this is the procedure we investigate, which we term the "surrogate-matrix" technique. This "surrogate" prefix seemed appropriate since the extra gauges are acting as a surrogate for the external loads. For brevity we will sometimes simply refer to gauge σ_{AC}, for example, which should be read as the gauge located on beam member AC that measures

the stress σ_{AC}. From now on the "matrix" technique is sometimes referred to as the "ordinary-matrix" technique to remind the reader that it is different to the surrogate-matrix technique.

Before carrying out the simulations let's, first reflect on possible outcomes. In the simple truss that we've chosen to simulate (see Figure 3.4), we decided to model what appeared to be the most ill-conditioned system for the chosen configuration (refer to § 3.4 for a more detailed discussion). We can have up to five strain gauges yielding different results on this truss, but three of those gauges are already being used in our investigation (remember we have simulated the relation $\sigma_{OC} = f(\sigma_{OA}, \sigma_{AB})$). Hence there are only two gauges available (σ_{AC} and σ_{BC}) to replace the two external loads (P_x and P_y)—which, by the way, is fortunate! If we use these two gauges, however, we run the risk of somehow improving the overall condition number of the modified system. This improved state follows since if we had already chosen the worst possible gauge configuration, then any surrogate gauges can only improve the situation. This whole argument would imply that any accurate solutions may be simply due to the fact that the modified system is well-conditioned (as compared to the ill-conditioned system used for simulations thus far).

We need to differentiate whether good results are; (i) due to (the possibility of) improved condition number (resulting from the surrogate good gauges); or (ii) due to the improved technique (namely the surrogate-matrix technique). The simulations were additionally carried out using the "good" gauges in order to reduce the effect of improving the condition number. In other words, we solved for the linear system $\sigma_{OC} = f(\sigma_{AC}, \sigma_{BC})$ (good gauges) using the ordinary-matrix technique, with the surrogate "bad" gauges (σ_{AC} and σ_{BC}) used for the surrogate-matrix technique. We use the terms "good" and "bad" gauges (or system) to refer to the condition number of the resulting system (of linear equations) determined using that configuration of gauges. These terms are relative, so that gauge configurations were judged to be good or bad depending on structural knowledge and simulation results. (For further details on gauge selection see § 3.4.)

Figure 4.9 shows how the surrogate-matrix technique performs. Both plots show normalised error versus approximation order (number of measurement points used to develop the solution). The top plot shows the bad system,

$$\sigma_{OC} = f\left(\sigma_{OA}, \sigma_{AB}; [P_x, P_y] \text{ or } [\sigma_{AC}, \sigma_{BC}]\right), \tag{4.63}$$

which in the surrogate-matrix solution replaces external loads (P_x and P_y) by the good gauges (σ_{AC} and σ_{BC}). Similarly, the bottom plot shows the good system,

$$\sigma_{OC} = f\big(\sigma_{AC}, \sigma_{BC}; [P_x, P_y] \text{ or } [\sigma_{AO}, \sigma_{AB}]\big), \tag{4.64}$$

which in the surrogate-matrix solution replaces external loads (P_x and P_y) by the bad gauges (σ_{OA} and σ_{AB}). These results add weight to the hypothesis that the surrogate-matrix technique performs just as well as the ordinary-matrix technique, even when the surrogate information is more ill-conditioned than the original system. If, on the contrary, the good results in the top plot were due to the improved condition number of the system, then the normalised error should be closer to that of the bottom plot, which it is not.

From Figure 4.9 we see that our suspicions about the ill-conditioning of the good (Equation (4.63)) and bad (Equation (4.64)) systems are confirmed. The normalised error

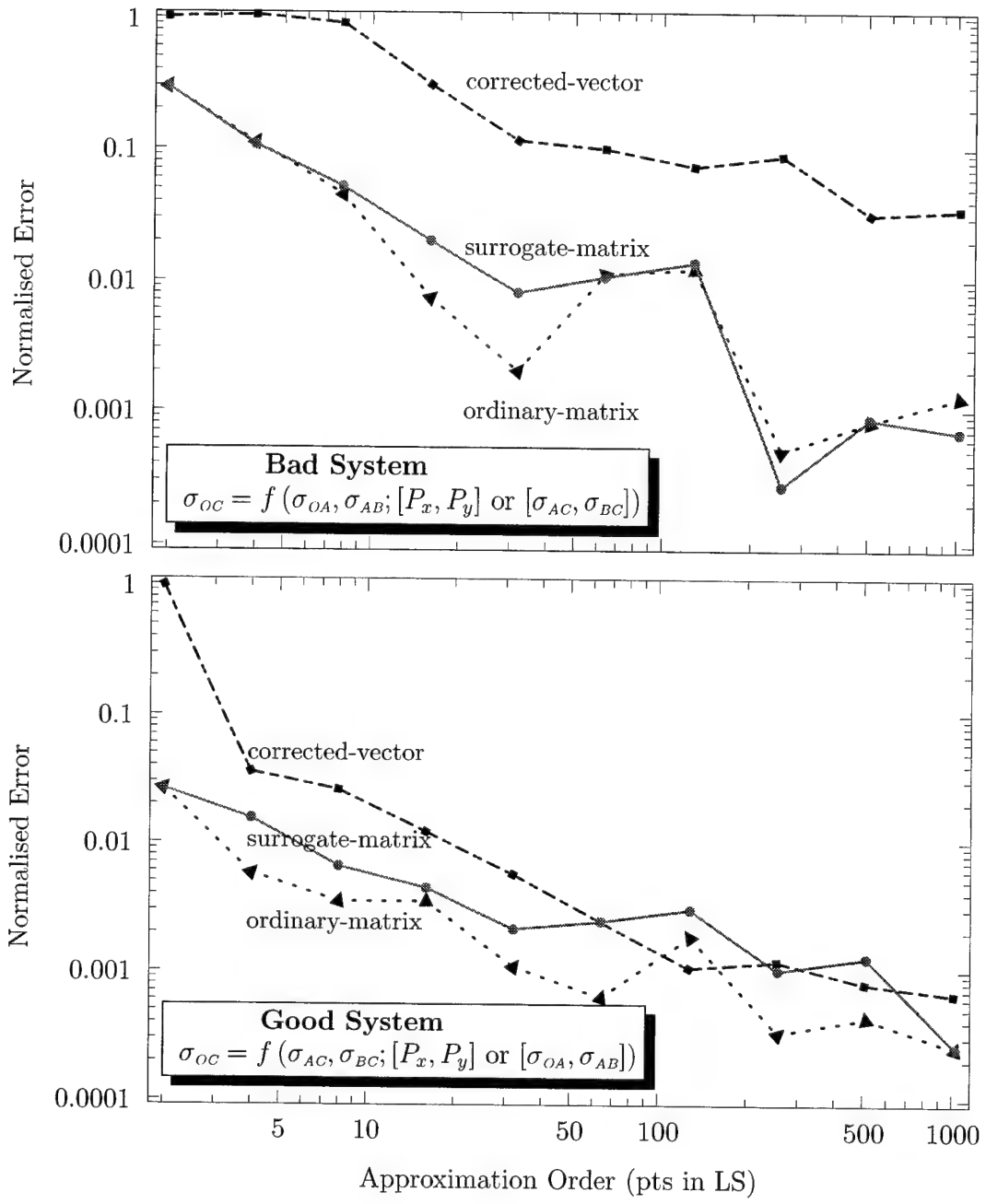


Figure 4.9: Comparison of the normalised error of the median solution for the corrected-vector (squares), ordinary-matrix (triangles), and surrogate-matrix (circles) techniques. The top plot shows the "bad" system, the system used for all simulations thus far. The bottom plot is the "good" system.

of the bad system (top plot) is, on the whole, worse than the normalised error of the good system (bottom plot). For our chosen truss configuration ($\alpha = 1.000$, $\beta = 2.000$, and

$\gamma = 1.100$, see page 16), the use of Equations (2.18) and (4.56) allows us to compute the bad system's condition number as $\kappa \approx 20.17$. A similar computation yields the condition number of the good system as $\kappa \approx 2.419$. For the chosen truss configuration (at the very least), these condition numbers confirm our suspicions about the relative good and bad assessment of the gauge configurations.

As a final point note that due to the improved condition number of the good system, the solutions of the ordinary-matrix and surrogate-matrix techniques weren't the only techniques to improve. All the vector-based techniques also improved, including the ordinary-vector technique. This emphasises the fact that for well-conditioned systems with small amounts of noise, there are only marginal gains to be made from using the matrix-based techniques over the vector-based techniques.

The surrogate-matrix technique performed as well as was theoretically predicted. Using additional truss gauges instead of external loads, the troublesome denominator found in the vector technique was successfully eliminated. We have simulation evidence to suggest that the similar accuracy between the ordinary-matrix and surrogate-matrix techniques is due to the underlying methods and not due to the improvement of the system's condition number. Of all the techniques sampled thus far, the surrogate-matrix has the highest accuracy, but requires additional gauges as compared to the vector-based techniques. We also noted that the vector-based techniques perform with comparable accuracy to the matrix-based techniques for well behaved systems.

4.11 Jack-Knife Correction and Bootstrap

We learned of the jack-knife correction (JKC) at the end of § 4.2 (see page 25), which improved the order of the approximating technique. Due to the vector technique's inconsistency, the JKC cannot be used on it. All the remaining techniques have the potential to be improved by the JKC; however, the variance of the solutions (see Figures 3.7 and 4.7) renders the JKC given by Equation (4.12) unusable. This follows since the large variance of solutions (for low order approximations) means that t_m in Equation (4.12) has a better than average chance of being far from the exact solution.

The JKC given by Equation (4.13) (the second order correction) doesn't suffer from this potentially bad initial estimate, and hence we suspect that it would provide a better correction than Equation (4.12) (the first order correction). The amount of computational effort required by the first and second order corrections (Equations (4.12) and (4.13) respectively) goes up linearly and quadratically, respectively, as the number of points used in the approximation.

To avoid the quadratic penalty of Equation (4.13), perhaps the re-formulation of Equation (4.12) as

$$t'_m = (m-1)\,\bar{t}_{m-1} - (m-2)\,\bar{t}_{m-2}$$

might provide a reasonable first order correction? In the above equation $\bar{t}_{m-2}$ should be computed using the original data set deleting two points at a time, but only applying this deletion $\mathcal{O}(m)$ times (since it is a first order correction). No analysis (neither experimental nor theoretical) of this suggested re-formulation was undertaken, and as such should not be used until verified.

The JKC uses the mean in its definition, see Equations (4.12) and (4.13). As already discussed this is not an appropriate measure for the multi-dimensional data at hand, and perhaps a good substitution would be the spatial-median.

Given that for each solution we are required to solve a least squares (LS) system (normally through the use of singular value decomposition), the JKC becomes computationally burdensome. However, since the development of the transfer function is essentially a once-off calibration, the large computational effort required may be warranted. Overall the JKC is recommended only if there is a lack of data.

In an earlier report [25] it was thought that the bootstrap method might improve the vector technique's results. We now realise that the errors in the vector technique were due to the technique's inconsistent nature, and hence the re-sampling offered by the bootstrap technique (see Press *et al.* [26] for example) would not eliminate these errors.

4.12 Two Input Surrogate Matrix Technique versus Four Input Redundant LS

In this section we determine if there is any advantage in using additional redundant information in the ordinary LS technique. We compare the surrogate matrix technique (with two gauges) to the redundant least squares (LS) technique (with four gauges). More specifically, the two input surrogate matrix technique has either one of the forms

$$\sigma_{OC} = A_1\sigma_{OA} + A_2\sigma_{AB} \qquad \text{or} \qquad \sigma_{OC} = B_1\sigma_{AC} + B_2\sigma_{BC} \tag{4.65}$$

depending on whether the good system or bad system respectively is being modelled, where A_1, A_2, B_1, and B_2 are constants. The four gauge redundant LS approximation has the form

$$\sigma_{OC} = C_1\sigma_{OA} + C_2\sigma_{AB} + C_3\sigma_{AC} + C_4\sigma_{BC}, \tag{4.66}$$

where C_1, C_2, C_3, and C_4 are constants. Notice that measurements from all four gauges are being used to predict the stress at gauge OA, despite the fact that only two gauges are needed to completely specify the system (see Equation (3.9) for example). We will show (for the one-dimensional case) that like the ordinary LS (that is, the vector) technique we have already investigated, the redundant LS technique is also inconsistent.

Let's begin with a simple one-dimensional linear system whose solution is given by

$$z = cx, \tag{4.67}$$

where z is completely specified by the constant c and variable x. Introduce a second variable y, which is a linear function of x, that is,

$$y = \theta x, \tag{4.68}$$

where θ is a constant. We can now write down z, given x and y, as

$$z = ax + by, \tag{4.69}$$

where both a and b are constants. The constant θ is a function of the three constants a, b, and c, and from the above three equations we have that $\theta = (c - a)/b$. Equation (4.69)

represents the one-dimensional equivalent of Equation (4.66); while Equation (4.67) is the one-dimensional ordinary LS solution of Equation (4.65). (Note that Equation (4.67) is the ordinary LS and not the surrogate matrix technique.)

We want to determine if there is any advantage in using the additional information given by y in developing our predictive system, given that this information (y) is correlated with information we already have (namely, x). In other words, we are using additional information that is redundant.

All the variables (x, y, and z) have noise, so we can use the results derived earlier. From Equation (4.16) we can write that

$$\frac{\widehat{a}}{a} - 1 = -\eta_x \frac{1 + \eta_y - r(b\eta_y\sqrt{s_{yy}})/(a\eta_x\sqrt{s_{xx}})}{(1+\eta_x)(1+\eta_y) - r^2}.$$

From Equation (4.68) we have that $r = 1$ (since x and y are completely correlated) and $s_{yy} = \theta^2 s_{xx}$, then the above expression becomes

$$\begin{aligned} \frac{\widehat{a}}{a} - 1 &= -\frac{1 + \eta_y - (b\eta_y|\theta|)/(a\eta_x)}{1 + \eta_y + \eta_y/\eta_x} \qquad (4.70) \\ &= -\frac{1 - \left|\frac{c-a}{b}\right| \frac{b}{a} \frac{\eta_y}{\eta_x(1+\eta_y)}}{1 + \frac{\eta_y}{\eta_x(1+\eta_y)}} \\ &= -\frac{1 - \Upsilon\Phi}{1 + \Phi}, \end{aligned}$$

where $\Upsilon = \text{sgn}(b/a)|c/a - 1|$ and $\Phi = \eta_y/[\eta_x(1+\eta_y)]$. (The operator "sgn" takes the sign of its argument, that is $\text{sgn}(x) = x/|x|$.)

We now see that, just like the ordinary LS approximation, the redundant LS approximation of a one-dimensional system (Equation (4.69)) is an inconsistent approximation of the true solution (except for the special case $\Upsilon\Phi = 1$). In other words, the error in Equation (4.69) will not tend to zero as the number of points used to develop the LS solution tends to infinity (unless $\Upsilon\Phi = 1$).

Let us now investigate the variance of the prediction error for the two approximations given by Equations (4.67) and (4.69). First define the relative error (or more pedantically, the negative of the relative error) in the coefficient a given by Equation (4.70) as

$$\begin{aligned} a_e &= 1 - \frac{\widehat{a}}{a} \\ &= \frac{1 + \eta_y - (b\eta_y|\theta|)/(a\eta_x)}{1 + \eta_y + \eta_y/\eta_x}, \end{aligned}$$

a similar expression arises for the coefficient b. Also define the prediction error of z as

$$\bar{e}_i = \bar{z}_i - z_i,$$

where $\bar{z}_i$ is the prediction given by Equation (4.69) and $z_i = ax_i + by_i$ is the exact solution (both for the ith measurement). The prediction $\bar{z}_i = \widehat{a}(x_i + u_i) + \widehat{b}(y_i + v_i)$ involves noise in both the x_i and y_i variables given respectively by u_i and v_i.

Assume the noise terms are uncorrelated with the both the signal and other noise terms (that is, $\epsilon = 0$ in the limit). After some tedious algebra the variance of the prediction error can be shown to be

$$\text{var}(\bar{e}) = (a\,a_e + \theta\,b\,b_e)^2 s_{xx} + a^2(1-a_e)^2 s_{uu} + b^2(1-b_e)^2 s_{vv}. \tag{4.71}$$

Re-arranging the relative error from Equation (4.70) gives that

$$a_e = \frac{\eta_x + \eta_x\eta_y - \eta_y|\theta|b/a}{\eta_x + \eta_x\eta_y + \eta_y}$$

and similarly for b_e we have that

$$b_e = \frac{\eta_y + \eta_x\eta_y - \eta_x a/(|\theta|b)}{\eta_x + \eta_x\eta_y + \eta_y}.$$

A small amount of algebra then gives the following three relations

$$a\,a_e + \theta\,b\,b_e = \frac{c\,\eta_x + [1-\text{sgn}(\theta)]\,[c - a(1-\eta_x/\eta_y)]}{1+\eta_x+\eta_x/\eta_y},$$

$$a(1-a_e) = \frac{c\,\text{sgn}(\theta) + a\,[1-\text{sgn}(\theta)]}{1+\eta_x+\eta_x/\eta_y},$$

and

$$b(1-b_e) = \frac{\eta_x}{\eta_y\,\theta}\,\frac{c - a\,[1-\text{sgn}(\theta)]}{1+\eta_x+\eta_x/\eta_y}.$$

To allow a comparison with the one-variable LS (Equation (4.67)) let's make the simplifying assumption that $\theta > 0$ (that is, there is a positive correlation between x and y). Substituting the above three simplified expressions into Equation (4.71) gives

$$\text{var}(\bar{e}) = \left(\frac{c}{1+\eta_x+\eta_x/\eta_y}\right)^2 \left\{\eta_x^2 s_{xx} + s_{uu} + s_{vv}\left[\frac{\eta_x}{\eta_y\,\theta}\right]^2\right\}.$$

Dividing both sides by s_{xx} (in other words $\text{var}(x)$) and using the relation

$$\frac{s_{vv}}{s_{xx}} = \frac{s_{vv}}{s_{yy}}\frac{s_{yy}}{s_{xx}} = \eta_y\theta^2$$

simplifies the above result to

$$\frac{\text{var}(\bar{e})}{\text{var}(x)} = c^2\left(\frac{\eta_x}{1+\eta_x+\eta_x/\eta_y}\right). \tag{4.72}$$

Now let's develop the variance of the prediction error for the one-variable approximation given by Equation (4.67). We have already developed the relative error of the coefficient c (see Equation (4.36)) from which we know that

$$\widehat{c} = c\,\frac{1}{1+\eta_x}.$$

Define the predictive error of the one-variable LS prediction (given by Equation (4.67)) as

$$\bar{\bar{e}}_i = \bar{\bar{z}}_i - z_i,$$

where $\bar{\bar{z}}_i = \widehat{c}(x_i + u_i)$. Following the same procedure as for the two-variable case we have that the variance of the predictive error for the one-variable approximation, given by Equation (4.67), is

$$\frac{\mathrm{var}(\bar{\bar{e}})}{\mathrm{var}(x)} = c^2 \left(\frac{\eta_x}{1 + \eta_x} \right). \tag{4.73}$$

Comparing the one-variable prediction (Equation (4.73)) and the two-variable prediction (Equation (4.72)) we see that the two-variable prediction produces superior predictions, especially when $\eta_x/\eta_y \gtrapprox 1$. Remember, however, that the one-variable ordinary LS prediction is inconsistent (as is the two-variable redundant LS technique).

A simulation of the five member truss (developed earlier) demonstrated the inconsistent behaviour of the redundant LS (Equation (4.66)) technique. Figure 4.10 shows plots of the predictive error for the surrogate matrix (given by Equation (4.65)) and the redundant LS (given by Equation (4.66)).

As before, we define the term "approximation order" as the number of simulation measurements used to develop the prediction equation (the horizontal axis in the plots of Figure 4.10). Unlike the previous simulation, we were not able to simply compare the approximate coefficients $\widehat{a}$ and $\widehat{b}$ to the exact coefficients a and b, since the redundant LS now contains four coefficients instead of two (see Equation (4.66)). As such we used the error in predicting the stress at gauge OC, as compared to the exact stress at gauge OC, as a measure of accuracy. For each approximation order we developed a set of 10 prediction equations (for both the surrogate matrix and redundant LS techniques). Each of these 10 equations were developed using different randomly generated measurement sets (which were statistically equivalent). We then randomly generated a new set of 100 measurements, and used this set to test the accuracy of the 10 prediction equations. In all, 1000 simulation measurements were used to compare the accuracy of both the surrogate matrix and redundant LS techniques for each approximation order.

The absolute values of the relative error (that is, $|(\widehat{\sigma}_{OC} - \sigma_{OC})/\sigma_{OC}|$) of these 1000 simulated measurements were then used to develop the spread of error plots shown in Figure 4.10. For both the surrogate matrix and the redundant LS shadings, the lower boundary represents the first quartile, the middle line the second quartile (or median), and the upper boundary the third quartile. The light grey and dark grey shadings represent the error spread for the surrogate matrix and redundant LS techniques respectively. The four gauge redundant LS plots for the "good" and "bad" systems would be identical if the same set of random data measurements were used, as it is these plots exhibit the same behaviour. Notice that as the approximation order increases the redundant LS plots tend to the same limit for both the "good" and "bad" systems.

The behaviour exhibited by the redundant LS is the same as that of the ordinary LS (compare Figure 3.7 and 4.10). For low order approximations the LS technique produces smaller errors than the surrogate matrix technique for the "bad" system (the results of the two techniques are comparable for the "good" system). However, for high order approximations the performance of the surrogate matrix technique is superior. Surprisingly,

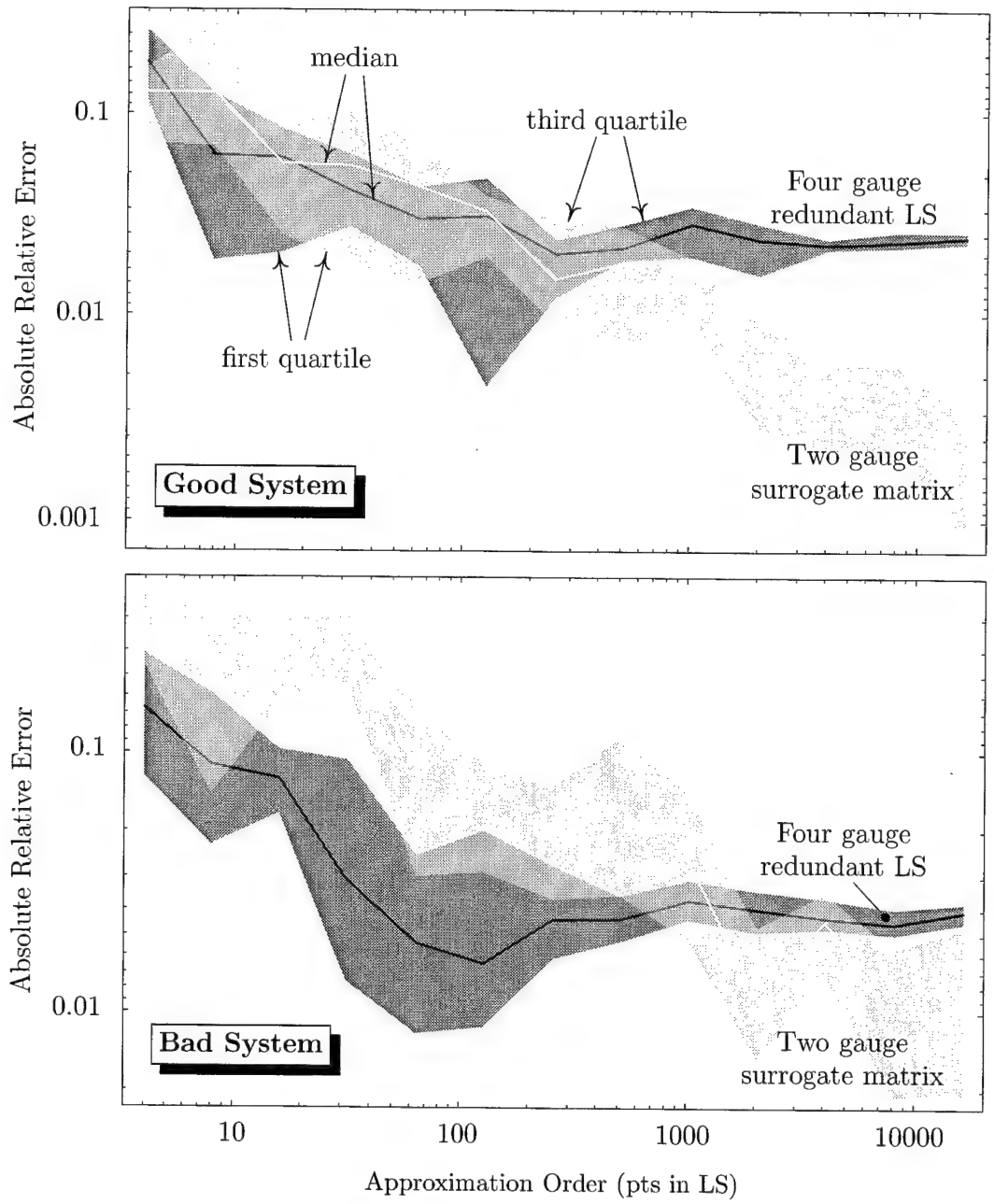


Figure 4.10: Plots of the predictive error spread for both the surrogate matrix (light grey) and the redundant LS (dark grey) techniques. The lower and upper boundaries of the shading represent the first and third quartiles respectively. The lines drawn near the centre of each shading represent the median (or second quartile).

the surrogate matrix technique even performed better for the "bad" system. This result is surprising since for the bad system (the bottom plot of Figure 4.10) we are only using the

two "bad" gauges in the surrogate matrix technique to predict the stress at the gauge OC. In contrast, the redundant LS technique uses all four gauges, including the two "good" gauges to estimate the stress at gauge OC.

Johnson and DiNardo [13, p. 88] provide results to a similar problem. They investigate the effects of redundancy (or collinearity) on the same system as Equation (4.69), except they assume there is no noise in the variables x and y. Under these assumptions they show that the variance of the constants $\widehat{a}$ and $\widehat{b}$ are

$$\text{var}(\widehat{a}) = \frac{s_{ww}}{(1-r^2)s_{xx}} \qquad \text{and} \qquad \text{var}(\widehat{b}) = \frac{s_{ww}}{(1-r^2)s_{yy}}$$

respectively, where s_{ww} is the amount of noise in the dependent variable z. These equations show that the variance of the coefficients $\widehat{a}$ and $\widehat{b}$ tend to infinity as the two inputs x and y become more correlated. It's interesting to note that under this collinear condition the results may be improved by increasing noise in the variables x and y. This noise increase would reduce the correlation between x and y, and hence (according to the above equations) reduce the variance of the $\widehat{a}$ and $\widehat{b}$ coefficients! We need to balance this unusual statement with results we have already established; namely, for noise in the variables x and y the LS technique is inconsistent. So that although the variance of the coefficient $\widehat{a}$ and $\widehat{b}$ reduces with increasing noise, the inconsistency increases. In other words, we are getting a tighter bound on a poorer solution.

In this subsection we have shown that the redundant LS technique is inconsistent (for the same reason as the ordinary LS technique). We also discovered that the redundant LS is marginally superior to the ordinary LS when noise is present, and leads to solutions with wide uncertainty intervals when the noise tends to zero.

5 Total Least Squares (TLS)

In the previous section we found that the ordinary least squares (LS) technique yielded inconsistent results when there was noise in both the input and output signals. The inconsistency was able to be corrected, resulting in the corrected LS technique. In this section we investigate the properties of another LS based technique termed total least squares (TLS), and determine which of these techniques is best suited to developing the transfer functions we require.

We begin by giving a geometric interpretation of the TLS. The next subsection reviews comparisons that have been made between the LS and TLS methods. We then investigate the variance of both the corrected LS and the total LS techniques. Finally, we report on simulations that confirm theoretical results presented in the preceding subsections.

5.1 Introduction to the Total Least Squares

Nievergelt [23] presents a short introduction into the TLS problem, where the derivation of the TLS solution is based on a geometric interpretation of results.

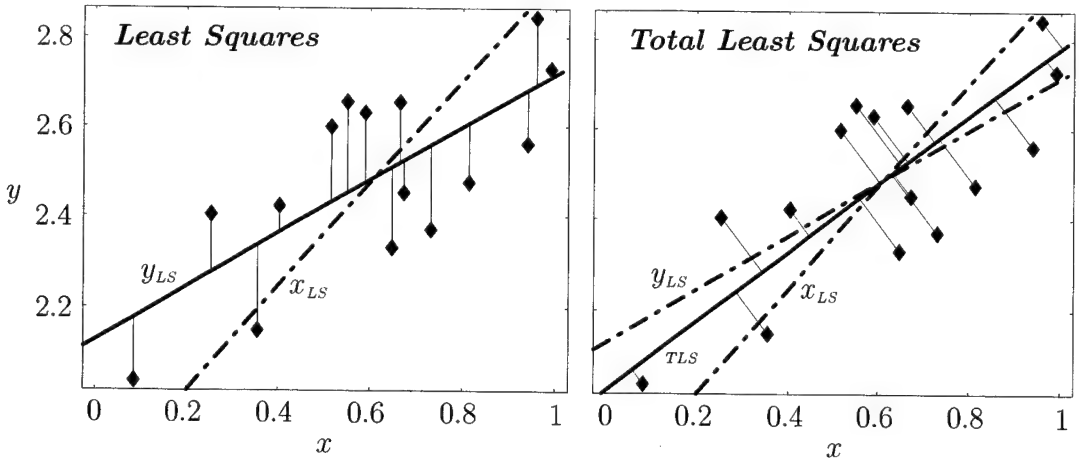


Figure 5.1: The lines x_{LS} and y_{LS} (left plot) minimise the distance to the data points in the x- and y- directions respectively. The right plot has the same data points and plots the TLS line, which minimises the perpendicular distance to the data points.

Figure 5.1 shows the difference between a LS solution and a TLS solution for two dimensional data. (A similar figure appears in Golub and van Loan [9] and also in Nievergelt [23].) The plot on the left shows the LS solutions x_{LS} (dashed line) and y_{LS} (solid line) that minimise the distance to the data points in the x- and y- directions respectively. (For clarity the horizontal distance from the data points to the x_{LS} line are not shown.) The plot on the right shows the TLS solution (solid line) that minimises the distance to the data points (independent of direction). From geometry we know that the shortest

distance from a point to a line lies in a perpendicular direction to the line passing through the desired point, and hence all distance emanating from the TLS line to the data points are perpendicular to the TLS line. For comparison the x_{LS} and y_{LS} lines are also shown in the right plot (drawn as dashed lines). Finally, notice that all three approximations (x_{LS}, y_{LS}, and TLS) pass through the centroid of the data points, and as might be expected the x_{LS} and y_{LS} solution bound the TLS solution.

The plots in Figure 5.1 highlight the fact that the TLS approximation is only superior if the variables used in the approximation are of equal importance. However, in our situation the output variable (or dependent variable) is more important than the input variable (or independent variable), and hence it is more important to minimise the error in the direction of the output variable. (Note that in some situations the error in some of the variables may be significantly larger than the error in the remaining variables. Under this condition it may be preferable to perform the LS in the direction of the variables with largest errors.)

As is clearly pointed out by Van Huffel and Vandewalle [35], the errors-in-variables model is useful when the primary goal is model parameter estimation rather than prediction. For prediction and also when the data significantly violates the model assumptions (for example, when outliers are present), the TLS estimates may be quite inferior to LS estimates.

For completion we list a result obtained by Golub and van Loan [9]: the condition of the TLS problem is always worse than the condition of the corresponding LS problem.

5.2 Comparing the LS and TLS Methods

In this subsection we begin our comparison by reviewing results related to the two-dimensional case.

Chen and Papadopoulos [5] compare the errors in approximating the slope and the z-intercept of the least squares (LS) and total least squares (TLS) lines. The exact linear system $z = ax + b$ is approximated by $\widehat{z} = \widehat{a}\widehat{x} + \widehat{b}$, where a and b are the slope and z-intercept of the exact line, and $\widehat{a}$ and $\widehat{b}$ are approximations of these constants. The true values of x and z have some error when measured and hence we obtain the values $\widehat{x}$ and $\widehat{z}$ respectively. Assuming the random noise is independent, has zero mean, and that the variance-covariance matrix exists, they develop errors for the slope and z-intercept using a first order series approximation. (They additionally assume, without loss of generality, that the m measured points used to develop the LS or TLS have zero mean.) The LS and TLS are found to have the same errors (to first order) for the slope and $\widehat{z}$-intercept given by

$$s_{\widehat{a}\widehat{a}} \approx \frac{a^2 s_{\widehat{x}\widehat{x}} + s_{\widehat{z}\widehat{z}} - 2a s_{\widehat{x}\widehat{z}}}{s_{xx}} \qquad \text{and} \qquad s_{\widehat{b}\widehat{b}} \approx \frac{a^2 s_{\widehat{x}\widehat{x}} + s_{\widehat{z}\widehat{z}} - 2a s_{\widehat{x}\widehat{z}}}{m},$$

where $s_{\widehat{a}\widehat{a}}$ and $s_{\widehat{b}\widehat{b}}$ are the variances of the LS or TLS slope and z-intercept respectively, $s_{\widehat{x}\widehat{x}}$ and $s_{\widehat{z}\widehat{z}}$ are the variances of the measured inputs $\widehat{x}$ and $\widehat{z}$ respectively, $s_{\widehat{x}\widehat{z}}$ is the covariance between the measurements, and s_{xx} is the variance of the exact x input. They conclude that to minimise the variance of the slope's error, the sample points x should be distributed as evenly as possible at the end points of the x input range (thus maximising s_{xx}).

Riggs, Guarnieri, and Addelman [27] empirically investigate thirty-four different fitting procedures (eleven previously published and twenty-three derived by the authors). They state that all good fitting procedures have three essential characteristics: invariance under linear transformations (the fitting coefficient is dimensionally correct), small root-mean square error, and a small bias. The empirical investigation involves approximating the line $z = x$ with the variable x varying over a range of 600 units. The exact x measurements were randomly and evenly generated using a Monte Carlo method and noise was added to the x_i and z_i values yielding $\widehat{x}_i = x_i + u_i$ and $\widehat{z}_i = z_i + w_i$ respectively. The variances of u and w (the amount of noise) were independently varied using one of 5.33%, 10.6%, 21.3%, or 42.7% noise. The number of points were also varied over 6, 12, 24, and 48.

Citing work from the 1800s and early 1900s Riggs *et al.* [27] provide the expression for the TLS approximation of a as

$$\widehat{a} = \frac{s_{\widehat{z}\widehat{z}} - s_{\widehat{x}\widehat{x}} + \sqrt{(s_{\widehat{z}\widehat{z}} - s_{\widehat{x}\widehat{x}})^2 + 4s_{\widehat{x}\widehat{z}}^2}}{2s_{\widehat{x}\widehat{z}}},$$

where $s_{\widehat{x}\widehat{x}} = \mathrm{var}(\widehat{x})$ and $s_{\widehat{x}\widehat{z}} = \mathrm{cov}(\widehat{x}, \widehat{z})$. The same expression shown above can be derived from Linnik [18, Eq. (0.1.29)], except the sign of the radical becomes $\pm$. The above equation is termed the "maximum likelihood" estimator by Anderson and Sawa [2].

Riggs *et al.* state that the TLS solution is variant under linear transformations (that is, incorrectly dimensioned), and cite several papers for the correctly dimensioned result

$$\widehat{a} = \frac{s_{\widehat{z}\widehat{z}} - \lambda s_{\widehat{x}\widehat{x}} + \sqrt{(s_{\widehat{z}\widehat{z}} - \lambda s_{\widehat{x}\widehat{x}})^2 + 4\lambda\, s_{\widehat{x}\widehat{z}}^2}}{2s_{\widehat{x}\widehat{z}}},$$

where $\lambda = s_{ww}/s_{uu}$ (that is, ratio of z noise to x noise). Johnston [12, Eq. (6-15)] obtains the result shown above, except the sign in front of the radical becomes $\pm$, and terms this the "classical errors-in-variables" approach. Johnston states that the sign in front of the radical is determined by the sign of the covariance between the input and output signal (that is, positive if $s_{\widehat{x}\widehat{z}} > 0$ and negative otherwise).

As the ratio of z noise to x noise tends to infinity (that is, $\lambda \to \infty$) the above expression becomes the classical LS technique (minimising the vertical distance between LS line and the measured points). Alternatively as the ratio of z noise to x noise tends to zero (that is, $\lambda \to 0$) the above expression becomes the LS technique for the regression of x on z (minimising the horizontal distance between LS line and the measured points). Johnston [12, p. 155] makes the same observations, and goes on to state that these two special cases ($\lambda \to 0$ and $\lambda \to \infty$) of the conventional LS solution may be regarded as extreme limiting cases of the general errors-in-variables model.

Due to the empirical nature of their work, Riggs *et al.* found it hard to draw definite conclusions.

5.3 Variance of the Corrected LS and the Total LS

In this subsection we compare the variance of both the corrected LS and the total least squares (TLS) solutions for a general n-dimensional system. In particular we review

results that compare the variance of these two techniques, and show that TLS always has a larger variance than LS (despite the fact that TLS is consistent and LS is not).

Assume the errors in all the data are independently and identically distributed with zero mean and the common variance matrix $\sigma^2 \boldsymbol{I}$ (where σ^2 is the error variance). Then Van Huffel and Vandewalle [35, p. 234] show that the TLS solution is a strongly consistent estimate of the parameters in the linear errors-in-variables model.

Consider the linear system given by Equation (4.41), and the related approximate linear system

$$\widehat{\boldsymbol{X}}\,\widehat{\boldsymbol{a}} = \widehat{\boldsymbol{z}},$$

where $\widehat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{U}$, $\widehat{\boldsymbol{z}} = \boldsymbol{z} + \boldsymbol{w}$, and $\boldsymbol{U}$ and $\boldsymbol{w}$ are the matrix and vector of input and output noise respectively . The closed-form expression for the basic TLS solution is given by [35, Eq. (2.27)]

$$\widehat{\boldsymbol{a}}_{TLS} = \left(\widehat{\boldsymbol{X}}^T \widehat{\boldsymbol{X}} - \sigma_{n+1}^2 \boldsymbol{I} \right)^{-1} \widehat{\boldsymbol{X}}^T \widehat{\boldsymbol{z}},$$

where σ_{n+1} is the smallest singular value of the augmented matrix $\left[\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{z}}\right] \in \mathbb{R}^{m \times (n+1)}$. The above solution, however, should not be used in the computation of the TLS solution. The generalised TLS problem (which considers a matrix $\widehat{\boldsymbol{Z}}$ of outputs instead of a vector $\widehat{\boldsymbol{z}}$) is given by Van Huffel and Vandewalle [34], who also provide a robust algorithm for the computation of the TLS solution (based on generalised SVD).

Fixed sample size distribution theory is termed "unwieldy" by Van Huffel and Vandewalle [34], and hence they consider only asymptotic distributions (that is, large sample sizes). Assume the errors in the input and output are independent and identically distributed with common zero mean vector and a common covariance matrix $(\sigma^2 \boldsymbol{I}_{n+1})$. Furthermore, assume that the covariance matrix of exact inputs

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \lim_{m \to \infty} \frac{1}{m} \boldsymbol{X}^T \boldsymbol{X}$$

exists, or in other words assume that there exists a linear solution to the system (4.41). Referring to Equation (4.44) we see that the above covariance matrix is the limiting case of the sample covariance matrix, that is, $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \lim_{m \to \infty} \boldsymbol{S}_{\boldsymbol{x}}$. The asymptotic distribution of $\sqrt{m}(\widehat{\boldsymbol{a}}_{TLS} - \boldsymbol{a})$ is then multivariate normal with zero means [34, p. 241].

Variance results are obtained if we further assume that $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is positive definite (that is, the underlying system $\boldsymbol{X}$ has full rank) and that the errors posses the third and fourth moments of a normal distribution. (Remember that a matrix has full rank if all the singular values are positive, see § 2.) Using these two additional assumptions, we know that $\sqrt{m}(\widehat{\boldsymbol{a}}_{TLS} - \boldsymbol{a})$ is asymptotically normally distributed with zero mean and covariance matrix [34, Eq. (8.45)]

$$\operatorname{var}_\infty \left[\sqrt{m} \left(\widehat{\boldsymbol{a}}_{TLS} - \boldsymbol{a} \right) \right] = \sigma^2 \left(1 + ||\boldsymbol{a}||_2^2 \right) \left[{\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{-1} + \sigma^2 {\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{-1} \left(\boldsymbol{I}_n + \boldsymbol{a}\boldsymbol{a}^T \right)^{-1} {\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{-1} \right], \quad (5.1)$$

which may be approximated as $\sigma^2(1 + ||\boldsymbol{a}||_2^2){\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{-1}$ for a small amount of noise (that is, $\sigma \ll 1$). The above expression differs from a result given by Schneeweiß [30] as we now show.

Define $\boldsymbol{\Sigma_u}$ as the covariance matrix of input errors

$$\begin{aligned}\boldsymbol{\Sigma_u} &= \lim_{m\to\infty} \frac{1}{m}\boldsymbol{U}^T\boldsymbol{U} \\ &= \text{diag}\left(\mathcal{E}(u_1)^2, \ldots, \mathcal{E}(u_n)^2\right),\end{aligned}$$

where u_j represents the noise on the jth input variable (which is the jth column of $\boldsymbol{U}$). The second line of the above displayed equation follows from the first line since we have assumed that the input noises are uncorrelated. Also define the *kurtosis matrix* as

$$\boldsymbol{K} = \text{diag}(k_1, \ldots, k_n), \qquad \text{where} \quad k_j = a_j^2\left\{\mathcal{E}\left(u_j^4\right) - 3\left[\mathcal{E}\left(u_j^2\right)\right]^2\right\} \quad \text{for } j = 1, \ldots, n$$

and a_j is the jth element of the coefficient vector $\boldsymbol{a}$. It is interesting to note that k_j/a_j^2 is exactly the definition of the fourth cumulant about the mean, see for example Kendall and Stuart [14, Eq. (3.43)]. The asymptotic variance of the corrected LS error is then given by [30, Eq. (5.4)]

$$\text{var}_\infty\left[\sqrt{m}\left(\widehat{\boldsymbol{a}}_{CLS} - \boldsymbol{a}\right)\right] = \boldsymbol{\Sigma_x}^{-1}\left\{\left[\mathcal{E}(z^2) + \boldsymbol{a}^T\boldsymbol{\Sigma_u}\boldsymbol{a}\right]\boldsymbol{\Sigma}_{\widehat{\boldsymbol{x}}} + \boldsymbol{K} + \boldsymbol{\Sigma_u}\,\boldsymbol{a}\,\boldsymbol{a}^T\boldsymbol{\Sigma_u}\right\}\boldsymbol{\Sigma_x}^{-1}, \quad (5.2)$$

where $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{x}}} = \lim_{m\to\infty}(1/m)\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}}$.

The TLS solution is simply a special case of the corrected LS solution [34, p. 243], in that for the TLS solution all noise terms are assumed to have the same magnitude (that is, $\mathcal{E}(z^2) = \sigma^2$ and $\boldsymbol{\Sigma_u} = \sigma^2\boldsymbol{I}$). Using this assumption together with the assumption that the fourth moment is the same as that of a normal distribution (that is, $\boldsymbol{K} = \boldsymbol{0}$), then Equation (5.2) becomes

$$\text{var}_\infty\left[\sqrt{m}\left(\widehat{\boldsymbol{a}}_{TLS} - \boldsymbol{a}\right)\right] = \sigma^2\left(1 + \boldsymbol{a}^T\boldsymbol{a}\right)\boldsymbol{\Sigma_x}^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{x}}}\boldsymbol{\Sigma_x}^{-1} + \sigma^4\boldsymbol{\Sigma_x}^{-1}\,\boldsymbol{a}\,\boldsymbol{a}^T\boldsymbol{\Sigma_x}^{-1}. \quad (5.3)$$

From the definition of $\widehat{\boldsymbol{X}}$ we have that $\widehat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{U}$ and hence

$$\frac{1}{m}\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}} = \frac{1}{m}\left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{X}^T\boldsymbol{U} + \boldsymbol{U}^T\boldsymbol{X} + \boldsymbol{U}^T\boldsymbol{U}\right).$$

Taking the limit as $m \to \infty$, remembering that we've assumed the input and noise are uncorrelated, we obtain

$$\begin{aligned}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{x}}} &= \boldsymbol{\Sigma_x} + \boldsymbol{\Sigma_u} \\ &= \boldsymbol{\Sigma_x} + \sigma^2\boldsymbol{I}_n. \end{aligned} \quad (5.4)$$

Equation (5.3) may then be written as

$$\text{var}_\infty\left[\sqrt{m}\left(\widehat{\boldsymbol{a}}_{TLS} - \boldsymbol{a}\right)\right] = \sigma^2\left(1 + ||\boldsymbol{a}||_2^2\right)\left[\boldsymbol{\Sigma_x}^{-1} + \sigma^2\boldsymbol{\Sigma_x}^{-1}\left(\boldsymbol{I}_n + \frac{\boldsymbol{a}\boldsymbol{a}^T}{1 + ||\boldsymbol{a}||_2^2}\right)\boldsymbol{\Sigma_x}^{-1}\right]. \quad (5.5)$$

Equations (5.1) and (5.5) are clearly different. The difference, however, is not as great as first appearances may suggest.

It can easily be shown that taking the product of the two expressions $\boldsymbol{I}_n \pm \boldsymbol{a}\boldsymbol{a}^T$ and $\boldsymbol{I}_n \mp \boldsymbol{a}\boldsymbol{a}^T/(1 + ||\boldsymbol{a}||_2^2)$ (note the opposite signs in front of the terms $\boldsymbol{a}\boldsymbol{a}^T$) gives the identity matrix (see Appendix A). We then have that

$$\boldsymbol{I}_n - \frac{1}{1 + ||\boldsymbol{a}||_2^2}\boldsymbol{a}\boldsymbol{a}^T = (\boldsymbol{I}_n + \boldsymbol{a}\boldsymbol{a}^T)^{-1},$$

where we've made use of the fact that $\boldsymbol{I}_n + \boldsymbol{a}\boldsymbol{a}^T$ is invertible (see Appendix A). Thus it appears as if there is a typographical error in one of the two equations for the covariance matrix of the TLS method (that is, Equation (5.1) or (5.5) derived by Van Huffel and Vandewalle [35] and Schneeweiß [30] respectively). Perhaps the sign in front of the term $\boldsymbol{a}\boldsymbol{a}^T$ should be reversed in one of these equations? At any rate, for small amounts of noise (that is, $\sigma \ll 1$) Equations (5.1) and (5.5) agree more closely (since the second term in both equations becomes small).

For completeness we also cite the variance of the corrected LS derived independently by Naidu [22] as $\text{var}_\infty(\widehat{\boldsymbol{a}}_{CLS} - \boldsymbol{a}) = (\sigma^2/m)(\boldsymbol{\Sigma}_{\widehat{\boldsymbol{x}}} - \boldsymbol{\Sigma}_{\boldsymbol{u}})^{-1}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{x}}}(\boldsymbol{\Sigma}_{\widehat{\boldsymbol{x}}} - \boldsymbol{\Sigma}_{\boldsymbol{u}})^{-1}$. Using the result given by Equation (5.4), this variance expression may be re-written as

$$\text{var}_\infty\left[\sqrt{m}\,(\widehat{\boldsymbol{a}}_{TLS} - \boldsymbol{a})\right] = \sigma^2{\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{-1} + \sigma^4{\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{-1}{\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{-1}, \tag{5.6}$$

which is different again from the two results already quoted in Equations (5.1) and (5.5). This difference, however, may be due to σ being defined (somewhat unclearly) as the norm of the residual (that is, if $\boldsymbol{\epsilon} = \widehat{\boldsymbol{X}}\,\widehat{\boldsymbol{a}} - \widehat{\boldsymbol{z}}$ then $\mathcal{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma^2\boldsymbol{I}_n$, it is unclear whether or not $\widehat{\boldsymbol{a}} = \widehat{\boldsymbol{a}}_{LS}$). Finally, a result with similar form to Equations (5.1), (5.5), and (5.6) is given by Fuller [7] in Theorem 2.3.2 and Equation 2.3.26, which are not included in this report due to the complexity in understanding the development.

Using Equation (5.1) the covariance of the TLS solution can be approximated by the matrix [34, Eq. (8.46)]

$$\text{cov}(\widehat{\boldsymbol{a}}_{TLS}) \approx \sigma^2(1 + ||\boldsymbol{a}||_2^2)(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \tag{5.7}$$

as compared to the covariance of the LS solution

$$\text{cov}(\widehat{\boldsymbol{a}}_{LS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

From the above two equations we see that the TLS solution always has a larger covariance than the LS solution. We have already come across this phenomenon for the univariate corrected LS case in § 4.5, see especially Equation (4.38).

For a finite sample size Equation (5.7) may be approximated by [34, Eq. (8.47)]

$$\text{cov}(\widehat{\boldsymbol{a}}_{TLS}) \approx \sigma^2(1 + ||\widehat{\boldsymbol{a}}||_2^2)({\widehat{\boldsymbol{X}}}^T\widehat{\boldsymbol{X}} - m\sigma^2\boldsymbol{I}_n)^{-1},$$

where $\sigma^2 \approx \sigma_{n+1}^2/m$ and σ_{n+1} is the $(n+1)$th singular value of the augmented matrix $[\widehat{\boldsymbol{X}}, \widehat{\boldsymbol{z}}]$. (Remember that m is the number of measurement samples.)

In this section we have shown that despite the LS procedure being inconsistent and the TLS being consistent, the variance of the TLS procedure is larger than that of the LS procedure. Furthermore, since the TLS is a special form of the corrected LS, we suspect that the variance of the corrected LS procedure is also larger than that of the LS procedure (a theory our truss simulations support).

5.4 Simulations Results

In this subsection we present some numerical simulations to support the theoretical work of the previous subsection.

Van Huffel and Vandewalle [34] performed Monte Carlo simulations and found that even for moderate sample sizes (20 observations or more) the asymptotic results detailed above provided good approximations. In their simulations they use the same measure as Ketellapper (see Equation (4.39)), namely the mean square error (MSE) given by

$$\begin{aligned}
\mathrm{MSE}\,(\widehat{\boldsymbol{a}}) &= \mathcal{E}\left[(\widehat{\boldsymbol{a}}-\boldsymbol{a})^T(\widehat{\boldsymbol{a}}-\boldsymbol{a})\right] \qquad (5.8)\\
&= \mathcal{E}\left\{[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})+\mathcal{E}(\widehat{\boldsymbol{a}})-\boldsymbol{a}]^T[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})+\mathcal{E}(\widehat{\boldsymbol{a}})-\boldsymbol{a}]\right\}\\
&= \mathcal{E}\left\{[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})]^T[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})]\right\}+\mathcal{E}\left\{[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})]^T[\mathcal{E}(\widehat{\boldsymbol{a}})-\boldsymbol{a}]\right\}\\
&\quad +\mathcal{E}\left\{[\mathcal{E}(\widehat{\boldsymbol{a}})-\boldsymbol{a}]^T[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})]\right\}+\mathcal{E}\left\{[\mathcal{E}(\widehat{\boldsymbol{a}})-\boldsymbol{a}]^T[\mathcal{E}(\widehat{\boldsymbol{a}})-\boldsymbol{a}]\right\}\\
&= \mathcal{E}\left\{[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})]^T[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})]\right\}+[\mathcal{E}(\widehat{\boldsymbol{a}})-\boldsymbol{a}]^T[\mathcal{E}(\widehat{\boldsymbol{a}})-\boldsymbol{a}] \qquad (5.9)\\
&= \text{total variance}(\widehat{\boldsymbol{a}})+\text{squared bias}(\widehat{\boldsymbol{a}}),
\end{aligned}$$

the cross-product terms being equal to zero since $\mathcal{E}[\widehat{\boldsymbol{a}}-\mathcal{E}(\widehat{\boldsymbol{a}})]=0$. (For a further discussion on the MSE measure see § 4.5 and also Kendall and Stuart [15].) The squared bias, total variance, and MSE of the simulation results were normalised by the Frobenius norm of the solution $||\boldsymbol{a}||_F^2$ and plotted on log-log scales. Van Huffel and Vandewalle then make the following observations.

- *TLS bias is much smaller than that of LS*, and both biases decrease as the number of observations m increases. This bias difference is more pronounced when the noise variance increases, m increases, $\boldsymbol{X}$ is ill-conditioned, or $\boldsymbol{a}$ is orientated towards the null space of $\boldsymbol{X}$.
- *TLS total variance is larger than that of LS.* The rank reduction of $\boldsymbol{X}$, due to large noise variances, causes an increase in bias but a decrease in variance of the solution.
- *TLS and LS solutions have comparable MSE for small noise variances, but for large noise variances the TLS performs better under the MSE measure.*

Van Huffel and Vandewalle cite other authors as giving similar, though more detailed, conclusions.

Heavy-tailed error distributions and outliers both deteriorate the performance of the TLS, where the clear dominance of TLS over LS is only apparent for large sample sizes (m large). These results are backed by both theoretical analysis and Monte Carlo simulations (see [35, p. 248] and citations therein).

6 Conclusion

We reviewed singular value decomposition (SVD) and determined when a two-by-two system would be both well- and ill-conditioned. We have seen that the SVD decomposition of a matrix is simply an orthogonalisation of the matrix into a sum of rank one matrices. The condition number of a two-by-two matrix allowed us to investigate the relation between gauge configuration and ill-conditioning. A two-by-two matrix becomes ill-conditioned when the determinate is zero, and is perfectly-conditioned when it has a symmetric-like form.

We then explained how the vector and matrix techniques worked. The development of a simple determinate truss allowed us to investigate the effects of gauge placement and measurement noise on the development of stress transfer functions. To obtain sensible stress solutions for a truss with two external loads, the minimum number of beam members required was found to be five. We developed the stress equations for each member of this truss, with beams of varying length (and hence varying truss geometries). We used this protean truss in simulations to support theoretical developments.

Assuming our predictive system was linear, we developed a relationship between the exact coefficients and the approximate coefficients. This assumption merely requires that the predictive system be a linear function of its variable, and not that the variables be linear in terms of the some global variables.

Further assuming that the noise is uncorrelated, we saw that as the two inputs become more correlated the troublesome denominator tends to zero, and hence the error tends to infinity. Thus our coefficient estimates (for the predictive system) deteriorate as the system becomes more ill-conditioned (even when only a small amount of noise is present).

To determine the behaviour of these coefficient estimations we further assumed that the two input signals and associated noises had the same variance. The error in the approximating coefficients is zero under two conditions: when the noise is zero or when the noise exactly cancels the correlation.

Assuming that the noise was much closer to zero than the correlation was to unity, we gained a greater understanding of the effects of exact coefficient ratio on errors. A contour plot of the noise amplifying function emphasised the importance of having input signals of a similar order of influence on the output signal. We found that if the exact coefficients had the same magnitude, then noise was no longer a problem, even for ill-conditioned systems. On the contrary, under this condition noise is beneficial. The well known result of least squares (LS) under-prediction (from errors-in-variables theory) was observed. From the contour plot we also discovered when the error in coefficient estimation was zero, and under what conditions the noise would be amplified detrimentally.

The LS technique (and hence the vector technique) was unbiased but inconsistent, and hence we developed a correction for the inconsistency. We developed conditions under which this LS correction might be sensitive to noise estimates, and showed that the use of bad noise estimations can easily send the correction to infinity. The superior performance of the matrix technique resulted from the elimination of the troublesome denominator (a function of correlation), making the matrix technique consistent.

We determined both the correlation and condition number of the five member truss

used for simulations. Correlated input signals do not imply an ill-conditioned system, and conversely, an ill-conditioned system does not imply correlated input signals. Due to the complexity in developing inequalities for near unit correlation and near ill-conditioning, we resorted to truss simulations. Unexpectedly, but rather fortunately, the simulations suggest that the correlation between input signals is always better than the system predicted ill-conditioning.

We have seen that the corrected-vector technique (based on corrected LS) is more accurate than the ordinary-vector technique (at least for large approximation orders). The accuracy, however, is still inferior to that of the matrix technique, which is most likely due to the sensitivity of errors in noise estimation.

The surrogate-matrix technique, which substituted additional gauges for external loads, performed as well as was theoretically predicted. Like the matrix technique, the surrogate-matrix technique eliminated the troublesome denominator found in the vector technique. Of all the techniques sampled thus far, the matrix-based techniques have the highest accuracy, but require additional gauges as compared to the vector-based techniques. We also noted that the vector-based techniques perform with comparable accuracy to the matrix-based techniques for well behaved systems.

Despite the LS procedure being inconsistent and the total least squares (TLS) being consistent, the variance of the TLS procedure is larger than that of the LS procedure. A review of the TLS technique demonstrated that for the development of stress transfer functions the corrected LS technique was better than the TLS, since the TLS was simply a special case of the corrected LS. These results explain why the variance of the corrected-vector technique was larger than the variance of the ordinary-vector technique.

In conclusion we recommend the use of the surrogate-matrix technique for ill-conditioned systems, due to its accuracy (in developing predictive linear models) and ease of implementation (no external loading is required). For well conditioned systems, we recommend the ordinary-vector technique, due to its ease of implementation (requires fewer gauges than the surrogate-matrix technique) and lower variance (as compared to the corrected-vector technique).

References

1. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, 9th edn, 1972.

2. T. W. Anderson and T. Sawa, Journal of the Royal Statistical Society. Series B, Statistical Methodology, 44 (1982), pp. 52–62.

3. A. H.-S. Ang and W. H. Tang, *Probability Concepts in Engineering Planning and Design: volume I, basic principles*, John Wiley, 1975.

4. K. E. Atkinson, *An Introduction to Numerical Analysis*, John Wiley, 2nd edn, 1989.

5. K.-W. Chen and A. S. Papadopoulos, *Comparison of the least squares and total least squares lines*, Metron, 54 (1996), pp. 93–101.

6. H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

7. W. A. Fuller, *Measurement Error Models*, Wiley, 1987.

8. R. N. Goldman and J. S. Weinberg, *Statistics, an Introduction*, Prentice-Hall, 1985.

9. G. H. Golub and C. F. van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

10. ——, *Matrix Computations*, Johns Hopkins University, 2nd edn, 1989.

11. R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge, 1985. (1996 Reprinting).

12. J. Johnston, *Econometric Methods*, McGraw-Hill, 1963.

13. J. Johnston and J. DiNardo, *Econometric Methods*, McGraw-Hill, 4th edn, 1997.

14. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics. Volume 1: Distribution Theory*, Griffin, 2nd edn, 1963.

15. ——, *The Advanced Theory of Statistics. Volume 2: Inference and Relationships*, Griffin, 3rd edn, 1973.

16. R. H. Ketellapper, *On estimating parameters in a simple linear errors-in-variables model*, Technometrics, 25 (1983), pp. 43–47.

17. A. M. Leahy, *Helicopter fatigue monitoring—the way ahead*, in Aerotech Conference, October 1993.

18. Y. V. Linnik, *Method of Least Squares and Priciples of the Theory of Observations*, Pergamon, 1961. (translated by R. C. Elandt, edited by N. L. Johnson).

19. E. Lloyd, ed., *Handbook of Applicable Mathematics. Volume VI: Statistics. Part A*, John Wiley, 1984.

20. D. C. LOMBARDO, *Helicopter structures—a review of loads, fatigue design techniques and usage monitoring*, Tech. Rep. 15, AR-00-137, Aeronautical and Maritime Research Laboratory, May 1993.

21. E. MARTIN, ed., *Mathematica 3.0 Standard Add-On Packages*, Wolfram Media, 1996.

22. L. K. NAIDU, Computational Statistics and Data Analysis, 10 (1990), pp. 143–151.

23. Y. NIEVERGELT, *Total least squares: State-of-the-art regression in numerical analysis*, SIAM Review, 36 (1994), pp. 258–264.

24. F. G. POLANCO, *Estimation of structural component loads in helicopters: A review of current methodologies*, Tech. Rep. DSTO-TN-0239, Aeronautical and Maritime Research Laboratory, December 1999.

25. ——, *Development of a stress transfer function for an idealised helicopter structure*, Tech. Rep. DSTO-RR-0171, Aeronautical and Maritime Research Laboratory, March 2000.

26. W. H. PRESS, S. A. TEUKOLSKY, W. T. VETTERLING, AND B. P. FLANNERY, *Numerical Recipes in C: the art of scientific computing*, Cambridge, 2nd edn, 1992.

27. D. S. RIGGS, J. A. GUARNIERI, AND S. ADDELMAN, *Fitting straight lines when both variables are subject to error*, Life Sciences, 22 (1978), pp. 1305–1360.

28. C. G. SCHAEFER, JR., *The effects of aerial combat on helicopter structural integrity*, in American Helicopter Society 45th Annual Conference, Boston, MA, 22–24 May 1989.

29. P. SCHMIDT, *Econometrics*, Dekker, 1976.

30. H. SCHNEEWEISS, *Consistent estimation of a regression with errors in the variables*, Metrika, 23 (1976), pp. 101–115.

31. C. G. SMALL, *A survey of multidimensional medians*, International Statistical Review, 58 (1990), pp. 263–277.

32. M. R. SPIEGEL, *Schaum's Outline Series: Theory and Problems of Statistics*, McGraw-Hill, 2nd edn, 1988.

33. G. W. STEWART, *Collinearity and least squares regression*, Statistical Science, 2 (1987), pp. 68–100.

34. S. VAN HUFFEL AND J. VANDEWALLE, *Analysis and properties of the generalized total least squares problem $AX \approx B$ when some or all columns in A are subject to error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 294–315.

35. ——, *The Total Least Squares Problem: computational aspects and analysis*, SIAM, 1991.

Appendix A: Properties of the Rank-One Matrix $\boldsymbol{a}\boldsymbol{a}^T$

In this appendix we derive several properties of the matrix $\boldsymbol{a}\boldsymbol{a}^T$, where $\boldsymbol{a}$ is a vector. In particular we show that $\boldsymbol{a}\boldsymbol{a}^T$ is idempotent-like and positive semidefinite. We finally use these results to prove the invertibility of the matrix $\boldsymbol{I} + \boldsymbol{a}\boldsymbol{a}^T$.

Consider the matrix $\boldsymbol{\Sigma_a} = \boldsymbol{a}\boldsymbol{a}^T$, where the vector $\boldsymbol{a} \in \mathbb{R}^{n\times 1}$ and hence the matrix $\boldsymbol{\Sigma_a} \in \mathbb{R}^{n\times n}$. Let the vector $\boldsymbol{a}$ have the elements a_i for $i = 1, \ldots, n$, then the matrix $\boldsymbol{\Sigma_a}$ has the form

$$\boldsymbol{\Sigma_a} = \begin{bmatrix} a_1^2 & a_1a_2 & \cdots & a_1a_n \\ a_2a_1 & a_2^2 & \cdots & a_2a_n \\ \vdots & \vdots & \ddots & \vdots \\ a_na_1 & a_na_2 & \cdots & a_n^2 \end{bmatrix}. \tag{A1}$$

We see that the matrix $\boldsymbol{\Sigma_a}$ must be symmetric and also rank one (or rank zero for $\boldsymbol{a} = \boldsymbol{0}$) since all rows are linearly dependent. Notice from the above equation that $\boldsymbol{\Sigma_a} \neq \boldsymbol{I}_n$ for any vector $\boldsymbol{a}$ since considering, for example, the first two elements in the equation $\boldsymbol{\Sigma_a} = \boldsymbol{I}_n$ we arrive at the contradiction $a_1 = a_2 = \pm 1$ but $a_1a_2 = 0$. Alternatively, we know that $\text{rank}(\boldsymbol{I}_n) = n$, and hence $\boldsymbol{\Sigma_a} \neq \boldsymbol{I}_n$.

We also see that if $||\boldsymbol{a}||_2 = 1$ then the matrix $\boldsymbol{\Sigma_a}$ is *idempotent* (that is $\boldsymbol{\Sigma_a}^2 = \boldsymbol{\Sigma_a}$). More generally we also have that

$$\begin{aligned} \boldsymbol{\Sigma_a}^2 &= \boldsymbol{a}\boldsymbol{a}^T\boldsymbol{a}\boldsymbol{a}^T \\ &= \boldsymbol{a}||\boldsymbol{a}||_2^2\boldsymbol{a}^T \\ &= ||\boldsymbol{a}||_2^2\boldsymbol{\Sigma_a}, \end{aligned} \tag{A2}$$

that is the square of the matrix $\boldsymbol{\Sigma_a}$ is a scaler multiple ($||\boldsymbol{a}||_2^2$) of itself. Johnston and DiNardo [13, p. 483] prove that each eigenvalue of an idempotent matrix is either zero or one. Using the above idempotent-like information we can easily generalise that proof to show that the eigenvalues of the matrix $\boldsymbol{\Sigma_a}$ are either zero or $||\boldsymbol{a}||_2^2$. We then know that the matrix $\boldsymbol{\Sigma_a}$ must be positive semidefinite (that is, all eigenvalues of $\boldsymbol{\Sigma_a}$ are non-negative).

Let λ and $\boldsymbol{z}$ be an eigenvalue and the corresponding eigenvector of the matrix $\boldsymbol{\Sigma_a}$ and let $\boldsymbol{B} = \boldsymbol{I}_n + \boldsymbol{\Sigma_a}$. By definition we then have that $\boldsymbol{\Sigma_a}\boldsymbol{z} = \lambda\boldsymbol{z}$, and upon adding $\boldsymbol{z}$ to both sides gives that

$$\begin{aligned} \boldsymbol{z} + \boldsymbol{\Sigma_a}\boldsymbol{z} &= \boldsymbol{z} + \lambda\boldsymbol{z} \\ (\boldsymbol{I}_n + \boldsymbol{\Sigma_a})\boldsymbol{z} &= (1 + \lambda)\boldsymbol{z}. \end{aligned}$$

(For a similar proof see [13, p. 482].) Remember, however, that $\boldsymbol{\Sigma_a}$ is positive semidefinite, and hence the matrix $\boldsymbol{B}$ must be positive definite (that is, all eigenvalues of $\boldsymbol{B}$ are positive). We know that the eigenvalues and singular values of a symmetric matrix are identical, and since $\boldsymbol{B}$ is positive definite it must also be invertible (refer to § 2 for an elucidation of these concepts).

Finally, let's consider the following product

$$\begin{aligned}
&(\boldsymbol{I}_n + \boldsymbol{\Sigma}_{\boldsymbol{a}}) \left(\boldsymbol{I}_n - \frac{1}{1 + ||\boldsymbol{a}||_2^2} \boldsymbol{\Sigma}_{\boldsymbol{a}} \right) \\
&\qquad = \boldsymbol{I}_n + \boldsymbol{\Sigma}_{\boldsymbol{a}} - \frac{1}{1 + ||\boldsymbol{a}||_2^2} \boldsymbol{\Sigma}_{\boldsymbol{a}} - \frac{1}{1 + ||\boldsymbol{a}||_2^2} \boldsymbol{\Sigma}_{\boldsymbol{a}}{}^2 \\
&\qquad = \boldsymbol{I}_n + \frac{1 + ||\boldsymbol{a}||_2^2 - 1}{1 + ||\boldsymbol{a}||_2^2} \boldsymbol{\Sigma}_{\boldsymbol{a}} - \frac{1}{1 + ||\boldsymbol{a}||_2^2} ||\boldsymbol{a}||_2^2 \boldsymbol{\Sigma}_{\boldsymbol{a}} \\
&\qquad = \boldsymbol{I}_n.
\end{aligned}$$

Since $\boldsymbol{I}_n + \boldsymbol{\Sigma}_{\boldsymbol{a}}$ is invertible we have that

$$\boldsymbol{I}_n - \frac{1}{1 + ||\boldsymbol{a}||_2^2} \boldsymbol{\Sigma}_{\boldsymbol{a}} = (\boldsymbol{I}_n + \boldsymbol{\Sigma}_{\boldsymbol{a}})^{-1}.$$

Index

DISTRIBUTION LIST

Effects of Noise on Almost Collinear Systems

Frank G. Polanco

	Number of Copies
AUSTRALIA	
DEFENCE ORGANISATION	
Task Sponsor	
S & T Program	
Chief Defence Scientist FAS Science Policy AS Science Corporate Management	1
Director General Science Policy Development	1
Counsellor Defence Science, London	Doc Data Sht
Counsellor Defence Science, Washington	Doc Data Sht
Scientific Adviser to MRDC, Thailand	Doc Data Sht
Scientific Adviser Policy and Command	1
Navy Scientific Adviser	Doc Data Sht
Scientific Adviser, Army	Doc Data Sht
Air Force Scientific Adviser	1
Director Trials	1
Aeronautical and Maritime Research Laboratory	
Director	1
Chief of Airframes and Engines Division	1
Research Leader Propulsion	1
Head of Helicopter Life Assessment (Ken F. Fraser)	1
Task Manager (Albert Wong)	1
Author (Frank Polanco)	1
Domenico C. Lombardo	1
Chris G. Knight	1
Soon-Aik Gan	1
Thomas Ryall	1
Shane Dunn	1
Hasan Shafi	1
Carl Mouser	1
Kate Lillingston	1
DSTO Libraries and Archives	
Library Fishermans Bend	Doc Data Sht

Library Maribyrnong	Doc Data Sht
Library Salisbury	1
Australian Archives	1
Library, MOD, Pyrmont	Doc Data Sht
US Defense Technical Information Center	2
UK Defence Research Information Centre	2
Canada Defence Scientific Information Service	1
NZ Defence Information Centre	1
National Library of Australia	1
Capability Systems Staff	
Director General Maritime Development	Doc Data Sht
Director General Land Development	1
Director General C3I Development	Doc Data Sht
Director General Aerospace Development	Doc Data Sht
Army	
ASNSO ABCA, Puckapunyal	4
SO(Science), DJFHQ(L), MILPO Enoggera, Qld 4051	Doc Data Sht
Commander Aviation Support Group, Oakey	1
Intelligence Program	
DGSTA Defence Intelligence Organisation	1
Manager, Information Centre, Defence Intelligence Organisation	1
Corporate Support Program	
Library Manager, DLS-Canberra	1
UNIVERSITIES AND COLLEGES	
Australian Defence Force Academy Library (ADFA)	1
Head of Aerospace and Mechanical Engineering, ADFA	1
Serials Section (M List), Deakin University Library, Geelong 3217	1
Hargrave Library, Monash University	Doc Data Sht
Librarian, Flinders University	1
OTHER ORGANISATIONS	
NASA (Canberra)	1
AusInfo	1

OUTSIDE AUSTRALIA

ABSTRACTING AND INFORMATION ORGANISATIONS

Library, Chemical Abstracts Reference Service	1
Engineering Societies Library, US	1
Materials Information, Cambridge Science Abstracts, US	1
Documents Librarian, The Center for Research Libraries, US	1

INFORMATION EXCHANGE AGREEMENT PARTNERS

Acquisitions Unit, Science Reference and Information Service, UK	1
Library – Exchange Desk, National Institute of Standards and Technology, US	1
Inderjit Chopra, Minta-Martin Professor and Director, Alfred Gessow Rotorcraft Center, Aerospace Engineering, University of Maryland, Maryland	1
Charlie Crawford, Chief Engineer, Aerospace and Transportation Laboratory, Georgia Tech Research Institute, Alabama	1
Prof Phil Irving, Head Damage Tolerance Group, School of Industrial and Manufacturing Science, Cranfield University, Cranfield	1
Dorothy Holford, Defence Evaluation and Research Agency, Farnborough, Hampshire	1

U.S. Army

Eric Robeson, Aviation Applied Technology Directorate, (Fort Eustis, Virginia)	1
Dr Wolf Elber, Director Vehicle Structures Directorate, NASA Langley Research Center (Hampton, Virginia)	1
Kevin Rotenberger, Aviation and Missile Command (Redstone Arsenal, Alabama)	1

U.S. Navy

Gene Barndt, Rotary Wing Structures, NAVAIRSYSCOM (Patuxent River, Maryland)	1

SPARES	3
Total number of copies:	60

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION DOCUMENT CONTROL DATA				1. CAVEAT/PRIVACY MARKING
2. TITLE Effects of Noise on Almost Collinear Systems			3. SECURITY CLASSIFICATION Document (U) Title (U) Abstract (U)	
4. AUTHOR Frank G. Polanco			5. CORPORATE AUTHOR Aeronautical and Maritime Research Laboratory 506 Lorimer St, Fishermans Bend, Victoria, Australia 3207	
6a. DSTO NUMBER DSTO–RR–0204	6b. AR NUMBER AR–011–785		6c. TYPE OF REPORT Research Report	7. DOCUMENT DATE March, 2001
8. FILE NUMBER M1/9/795	9. TASK NUMBER DST 98/210	10. SPONSOR CDS	11. No OF PAGES 77	12. No OF REFS 35
13. URL OF ELECTRONIC VERSION http://www.dsto.defence.gov.au/corporate/reports/DSTO–RR–0204.pdf			14. RELEASE AUTHORITY Chief, Airframes and Engines Division	

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved For Public Release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, SALISBURY, SOUTH AUSTRALIA 5108

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS

No Limitations

18. DEFTEST DESCRIPTORS

military helicopters, rotors, fatigue (materials), stress measurement

19. ABSTRACT

We investigate the effects of noise on developing predictions of ill-conditioned systems from measurements. In particular we investigate collinearity between measurement devices. We assume the system is linear in the measurements taken, and that the measurement noise is uncorrelated both to the true measurements and to other measurement noises. The "matrix" and "vector" techniques (two stress transfer function techniques developed earlier) are analysed. The matrix technique produces better results, but requires external system information during calibration. On the other hand, the vector technique (based on least squares) is easily implemented (no knowledge of external information is required), but is sensitive to ill-conditioned configurations of measuring devices. The vector technique is shown to be the well-known errors-in-variable model, and hence unbiased but inconsistent, which explains the large errors it produces. Although a correction to the vector technique improves results, it is still not as accurate as the matrix technique. This vector correction additionally requires estimates of noise in the measuring devices, and suffers from sensitivity to noise estimation errors. The surrogate-matrix technique substitutes internal for external system information, circumventing the need for external system measurements. Simulation results involving a simple truss support all theoretical findings. The surrogate-matrix and vector techniques are recommended for ill- and well-conditioned systems respectively.